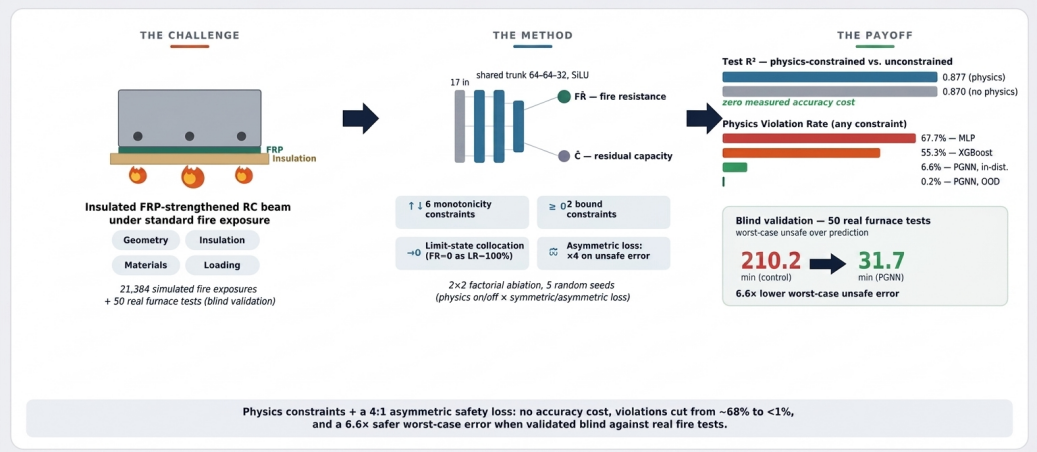
Graphical Abstract

Physics-guided neural network with an asymmetric safety loss for fire resistance prediction of insulated FRP-strengthened RC beams

Nripesh Poudel, Sadiksha Kandel



Highlights

Physics-guided neural network with an asymmetric safety loss for fire resistance prediction of insulated FRP-strengthened RC beams

Nripesh Poudel, Sadiksha Kandel

- Enforces six monotonicity, two bound, and one limit-state constraint for fire resistance surrogates.
- Asymmetric loss penalises unsafe over-predictions 4× more heavily than under-predictions.
- Physics Violation Rate drops from 67.7% (MLP) and 55.3% (XGBoost) to 6.6%.
- Constraints cut out-of-distribution safety loss dispersion sevenfold.
- Blind validation on 50 real tests confirms conservative error transfer.

Physics-guided neural network with an asymmetric safety loss for fire resistance prediction of insulated FRP-strengthened RC beams

Nripesh Poudel[a,*], Sadiksha Kandel[a]

[a]*Department of Civil Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Pulchowk, Lalitpur, 44700, Bagmati Province, Nepal*

Abstract

Machine-learned surrogates predict fire resistance (FR) of insulated FRP-strengthened RC beams accurately but violate basic physics, extrapolate unpredictably, and treat unsafe over-predictions the same as conservative under-predictions. We develop a physics-guided neural network (PGNN) that enforces six monotonicity constraints via automatic differentiation, two bound constraints, and a limit-state collocation constraint beyond the data envelope, trained with an asymmetric loss that penalises over-prediction four-fold. On 21,386 simulated fire exposures, a 2×2 factorial ablation over five seeds shows physics constraints cost no accuracy (test $R^2 = 0.877 \pm 0.005$ vs 0.870 ± 0.014 unconstrained) while a model-agnostic Physics Violation Rate—introduced here for fair comparison with tree ensembles—drops from 67.7% (unconstrained MLP) and 55.3% (XGBoost) to 6.6%, and to 0.2% out-of-distribution. The asymmetric loss converts an implicit design margin into a learned one, cutting unsafe-error rate by 40% in-distribution and over 60% out-of-distribution. Physics and asymmetric constraints also reduce seed-to-seed dispersion of out-of-distribution safety behaviour sevenfold—a property we term constraint-induced reliability. Blind validation on 50 real furnace tests, graded by extrapolation severity, confirms the conservative failure mode transfers to real beams (with deployment guard, worst-case unsafe error 31.7 min vs 210.2 min for unconstrained control) while revealing that narrow material-property ranges in the simulation campaign prevent accu-

*Corresponding author.
Email address: 079msste015.nripesh@pcampus.edu.np (Nripesh Poudel)

racy validation of any surrogate—identifying training-envelope widening as the binding constraint.

Keywords: fire resistance, FRP-strengthened beams, physics-guided neural network, monotonicity constraints, asymmetric loss, surrogate modelling

---

1. Introduction

1.1. Background and Motivation

Externally bonded fibre-reinforced polymer (FRP) strengthening is widely adopted for reinforced concrete (RC) beams due to high strength-to-weight ratio, corrosion resistance, and fast installation [1, 2, 3].

But FRP systems are vulnerable to fire. The adhesive layer degrades rapidly as it approaches the glass transition temperature (T_g), typically 60 to 120 °C—well below temperatures reached in even minor compartment fires. After fire, strengthening effect drops sharply, and the member must rely on its unstrengthened capacity plus whatever protection the insulation provides [4].

Fire resistance (FR) of an insulated FRP-strengthened beam is the time to flexural failure under standard fire exposure. It depends nonlinearly on geometry, insulation, thermal properties, material properties, and sustained load.

For fire-resistance design, assessment methods sit at opposite ends of the cost spectrum. Standard furnace tests give direct experimental evidence but are expensive and time-consuming, resulting in limited databases for FRP-strengthened members [5, 6]. Validated numerical models using coupled heat-transfer and structural analyses can account for temperature-dependent degradation of concrete, steel, FRP, insulation, and bond, reproducing observed thermal and structural responses with reasonable accuracy [7, 8]. But high-fidelity fire simulations remain computationally intensive, limiting their use in iterative applications like design optimisation, parametric studies, uncertainty quantification, and reliability analysis [9, 10]. This cost asymmetry motivates machine-learning surrogates: after training on a database of experimental or numerical results, a surrogate provides rapid predictions for design queries at a fraction of the simulation cost [11, 12].

1.2. Machine learning surrogates in structural fire engineering

Data-driven prediction of fire resistance and elevated-temperature response has grown rapidly. Ensemble trees (random forests, gradient boost-

ing, XGBoost) and multilayer perceptrons applied to RC columns, beams, slabs, steel members, and composite systems routinely report $R^2 > 0.9$ on held-out data [13, 14]. Recent studies show machine learning models achieve high accuracy for fire-resistance time and residual capacity of FRP-strengthened members, though typically investigated separately rather than in a unified framework [14, 15, 16]. Interpretability techniques (feature importance, partial dependence, SHAP) extract engineering insight from these models [12, 17].

Three limitations are widely documented in surrogate-modelling and physics-guided machine-learning literature but remain largely unaddressed in structural fire applications [18, 19, 14, 20].

First, physical inconsistency. A purely data-driven model fitted to finite data need not satisfy known physical relationships and may predict, for example, fire resistance dropping as insulation thickness increases, or rising as applied load increases [18, 19, 21]. Such trends contradict established thermal and structural behaviour: increased insulation delays temperature rise, while increased load reduces resistance under fire [7, 8, 12]. These violations are not cosmetic. When a model investigates design sensitivities—say, whether thicker insulation improves fire resistance—a wrong-signed response fails the intended use, even when aggregate error metrics look satisfactory.

Second, uncontrolled extrapolation. Behaviour outside training data is not determined by interpolation accuracy alone and depends strongly on model class and inductive assumptions [22]. Regression forests predict via averages of training responses in terminal leaves and stay bounded by observed values, often saturating during extrapolation [23, 24]. Neural networks extrapolate beyond observed target ranges, but without physical constraints their responses outside the training domain need not follow physical trends and may be unreliable [22]. This matters because design queries often extend beyond parameter combinations in any finite database.

Third, risk-neutral training. Mean-squared error is symmetric, assigning equal penalty to positive and negative errors of equal magnitude. A 20-min over-prediction of fire resistance—potentially unsafe—is penalised the same as a 20-min under-prediction, which is conservative. Symmetric objectives ignore directional consequences when consequences are unequal [25]. For fire resistance, over-prediction and under-prediction do not have equivalent safety implications.

1.3. Physics-informed and physics-guided learning

Physics-informed machine learning addresses the first two deficiencies. Since PINNs [26], models that embed governing equations or known constraints into the training objective have spread across computational mechanics, with applications in structural and fire engineering [27, 28]. The dominant paradigm uses differential-equation residuals, which works when the PDE is known and tractable. For FRP-strengthened beams in fire, the domain knowledge is different: monotonicity relations (FR non-decreasing in insulation thickness, depth, and T_g; non-increasing in conductivity, load, and load ratio); bound constraints (FR ≥ 0); residual capacity bounded by ambient capacity; and limit-state behaviour (FR collapsing to zero as load ratio approaches 100% of ambient capacity). Enforcing inequality and monotonicity constraints requires gradient-penalty and collocation mechanisms, distinct from PDE-residual methods [26, 29, 21, 30]. The limit-state law governs behaviour beyond any available data, so it cannot be learned from data alone and must be imposed through synthetic collocation points.

Three gaps remain in the physics-guided literature. Evaluation relies on symmetric accuracy metrics that hide the safety direction of errors and misrepresent conservative estimators. Ablation studies rarely isolate individual mechanisms (constraints versus modified objectives). Results are overwhelmingly single-seed, leaving unclear whether reported safety behaviour is reproducible or an artifact of initialization. This matters for certification: a model class whose extrapolation varies wildly between training runs cannot be certified by training just one member.

1.4. Problem statement and research gap

Fast surrogate models for the fire resistance of insulated FRP-strengthened RC beams achieve high accuracy within the training distribution, but violate fundamental physics for most queries. They extrapolate arbitrarily outside their training experience and treat unsafe and conservative errors as equivalent. Standard practice does not detect these failures.

To our knowledge, no prior work on FRP-strengthened beams under fire combines:

- monotonicity constraints enforced through automatic differentiation of the network's input–output map;
- bound and limit-state constraints enforced on synthetic collocation points beyond the data envelope;

- an asymmetric, risk-weighted training objective encoding the safety asymmetry of the prediction task;
- a factorial ablation separating each mechanism's contribution;
- multi-seed replication of both accuracy and safety metrics;
- a blind experimental validation against real furnace tests, analysed as a graded sim-to-real stress test rather than an undifferentiated holdout.

1.5. Objectives and Contributions

The objectives are:

1. Develop a physics-guided neural network (PGNN) for fire resistance of insulated FRP-strengthened RC beams that enforces six monotonicity constraints, two bound constraints, and a limit-state collocation constraint within a single training objective, alongside an asymmetric safety loss that penalises over-prediction four times as heavily as under-prediction.
2. Quantify the individual and joint effects of physics constraints and the asymmetric objective on accuracy, error direction, and physical consistency, through a 2×2 factorial ablation replicated over five random seeds, benchmarked against XGBoost, LightGBM, random forest, and MLP baselines.
3. Introduce and apply a model-agnostic Physics Violation Rate (PVR), a finite-perturbation measure of physical consistency applicable identically to tree ensembles and neural networks, on both in-distribution and out-of-distribution data.
4. Validate the framework blind against 50 real furnace tests, grading each test by extrapolation severity relative to the training envelope and evaluating raw and deployment-guarded predictions with a directional, safety-oriented metric suite.

The study makes four contributions. First, physics constraints do not cost predictive accuracy while cutting physics violations from 55–68% to under 7% in-distribution and 0.2% out-of-distribution. Second, joint imposition of physics and asymmetric constraints shrinks seed-to-seed variability of out-of-distribution safety behaviour roughly sevenfold — a property we call constraint-induced reliability. Third, the engineered conservative failure

direction transfers from simulation to real fire tests, even under severe covariate shift. Fourth, the training-envelope gap — simulation campaigns using material-property ranges far narrower than actual construction — is the main obstacle preventing experimental validation of any surrogate architecture for this problem.

The remainder of the paper is organised as follows. Section 2 details the datasets, the PGNN formulation, the ablation and replication design, the PVR methodology, and the blind validation protocol. Section 3 presents the results together with their discussion. Section 4 concludes and outlines future work.

2. Methodology

2.1. Overview and Design philosophy

The methodology comprises five sequential phases: data preparation with a deliberate out-of-distribution holdout; training of conventional baselines; design and training of physics-guided neural networks under a factorial ablation with multi-seed replication; quantification of physical consistency through a model-agnostic Physics Violation Rate; and blind validation against real furnace tests, analysed as a graded sim-to-real stress test. Workflow is represented in Figure 1. The design is governed by two principles throughout. The first principle is unconfounding: every mechanism that the proposed model adds is isolated within an ablation cell, so its individual contribution to accuracy, safety, and physical consistency is not merely asserted but measurable. The second principle is directional evaluation, because the target quantity is safety-critical. All evaluations report the direction and magnitude of errors, not merely their symmetric magnitude, using a fixed metric suite applied identically to every model regardless of its training objective.

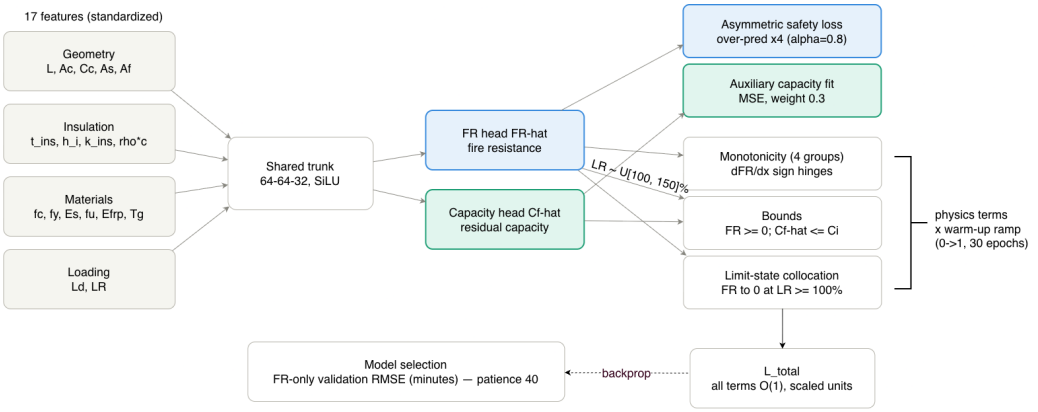


Figure 1: Workflow schematic of the five-phase methodology, showing the data flow from the 21,386-record simulation corpus and the 50-beam reserved furnace-test set through model training, ablation, physical-consistency auditing, and blind validation. The 17 standardized input features are organized into four physically meaningful groups—Geometry, Insulation, Materials, and Loading—and passed through a shared three-layer trunk (64–64–32 units, SiLU activation) with two linear output heads for fire resistance, $\hat{FR}$, and auxiliary residual capacity, $\hat{C}$. The training objective combines the asymmetric safety loss, auxiliary capacity-fit loss, grouped monotonicity constraints, bound constraints, and limit-state collocation constraint. Physics terms are scaled and introduced using a 30-epoch linear warm-up, while model selection is based solely on validation fire-resistance RMSE with early stopping (patience = 40). Detailed definitions of the loss terms and constraints are provided in Sections 2.5.2–2.5.7.

2.2. Data sources and preparation

2.2.1. Simulation corpus

The primary corpus consists of 21,386 numerically simulated fire exposures of insulated FRP-strengthened reinforced concrete beams, generated using a thermo-mechanical finite-element model [6]. Each record describes one beam–insulation–load configuration exposed to a standard fire and carries the resulting fire resistance, in minutes — the time to flexural failure — as the prediction target. The source sheet contains 21,399 data rows; thirteen were removed during loading — six entirely blank, and seven with missing or unparseable entries in a feature, target, or auxiliary column — leaving the 21,386 records analysed throughout. Two auxiliary quantities, initial flexural capacity and final flexural capacity, were retained per record. These are not prediction targets themselves but supply ground truth for one of the physics constraints.

One source-corpus artefact is disclosed here and retained unaltered. A

total of 793 records, 3.7% of the corpus, carry a small negative nominal insulation thickness, with a minimum of -2.03 mm (Table 1). Such values are not physically realisable and evidently arise in dataset generation rather than describing any beam that could be built. They are retained because excising them would change the corpus on which every result was computed; their practical consequence is that the t_{ins} monotonicity constraint is enforced over a range whose lower tail extends marginally below zero, which does not affect the sign of the constraint. Regenerating the corpus under a non-negative thickness bound is recommended alongside the envelope widening identified in 3.9.

2.2.2. Feature space

Each beam is described by 17 features across four physically meaningful groups:

- Geometry (5): span L (mm); concrete cross-sectional area A_c (mm^2); concrete cover C_c (mm); tension steel area A_s (mm^2); FRP reinforcement area A_f (mm^2).
- Insulation (4): insulation thickness t_{ins} (mm); insulation depth on the web h_i (mm); insulation thermal conductivity k_{ins} ($\text{W m}^{-1}\text{K}^{-1}$); volumetric heat capacity ρc ($\text{J m}^{-3}\text{K}^{-1}$).
- Materials (6): concrete compressive strength f_c (MPa); steel yield strength f_y (MPa); steel modulus E_s (MPa); FRP tensile strength f_u (MPa); FRP modulus E_{frp} (MPa); adhesive glass transition temperature T_g (°C).
- Loading (2): applied load L_d (kN); load ratio LR (%), the applied moment as a percentage of ambient flexural capacity.

All features with statistics are represented in Table 1.

Table 1: Descriptive statistics of the input features and target for the simulation corpus (n=21 386) and the real fire test set (n=50). The disparity between simulated and real material-property ranges (f_y, f_c, E_s) anticipates the training-envelope gap analysed in the experimental validation. The negative minimum of t_{ins} is a source-corpus artefact, discussed in 2.2.1.

			Simulation corpus				Real fire tests			
Group	Feature	Unit	Mean	Std	Min	Max	Mean	Std	Min	Max
Geometry	L	mm	3,363	1,466	1,220	6,200	2,864	1,239	1,260	6,000
	A_c	mm^2	88,493	41,611	12,000	157,500	55,170	38,616	12,000	125,730
	C_c	mm	26.5	9.72	10.0	38.0	21.1	8.23	10.0	38.0
	A_s	mm^2	400.7	205.6	63.3	854.6	319.9	279.6	56.5	942.5
	A_f	mm^2	118.2	103.2	0.00	629.3	67.8	72.7	0.00	460.0
Insulation	t_{ins}	mm	25.3	10.5	−2.03	38.0	21.7	16.4	0.00	50.0
	h_i	mm	41.6	35.3	0.00	114.0	119.7	158.2	0.00	500.0
	k_{ins}	W m^{-1}K^{-1}	0.11	0.06	0.00	0.23	0.12	0.15	0.00	0.67
	ρc	J m^{-3}K^{-1}	453,015	241,503	0.00	1,031,000	441,291	325,171	0.00	1,320,000
Materials	f_c	MPa	33.5	4.06	27.0	40.0	36.7	8.11	23.0	52.0
	f_y	MPa	436.6	13.6	414.0	460.0	503.2	69.3	364.0	591.0
	E_s	MPa	204,986	4,090	200,000	210,000	204,400	6,734	193,000	210,000
	f_u	MPa	2,116	1,557	0.00	4,840	2,125	1,179	0.00	4,030
	E_{frp}	MPa	134,203	78,695	0.00	255,530	143,734	72,775	0.00	260,000
	T_g	°C	81.3	10.1	0.00	88.0	56.5	26.4	0.00	85.0
Loading	L_d	kN	58.2	31.3	3.82	220.0	58.9	37.9	7.20	140.0
	LR	%	56.3	11.5	15.3	73.5	43.1	10.1	30.0	66.0
Target	FR	min	101.9	73.9	5.00	300.0	111.9	55.2	12.0	240.0

2.2.3. Reserved experimental set

A second dataset, comprising 50 real furnace tests on insulated FRP-strengthened beams, was compiled and reserved from the outset for blind validation [6]. It was not used for training, scaling, model selection, hyperparameter tuning, or threshold setting in any phase. The dataset required unit harmonisation prior to use. Spans recorded in metres were converted to millimetres, and load ratios recorded as fractions were converted to percentages. One physically impossible FRP-modulus entry (958 GPa) was corrected against magnitude plausibility. An automated guard was also provisioned, which re-checks the conversions at every run and reports zero corrections against the harmonised workbook.

2.3. Partitioning: in-distribution splits and the OOD holdout

A central design choice was made before any random splitting: an out-of-distribution (OOD) extrapolation set was carved out. The holdout rule reserves every record above the 95th percentile of either insulation thickness or load ratio. In this corpus the two conditions are not symmetric. The

95th percentile of t_{ins} coincides with the corpus maximum of 38.0 mm, so the thickness condition selects no records at all, and all 1,058 held-out records (4.95% of the corpus) are selected by the load-ratio condition, LR $> 71.2\%$. The holdout is therefore in effect a load-ratio tail holdout, and it is described as such throughout; we report it this way rather than as a joint thickness-and-load-ratio extrapolation, which the data would not support. These records represent the load extremes of the campaign, meaning that if a model fails on them, it will also fail on realistic queries rather than only on exotic ones. The remaining 20,328 in-distribution records are split in a 70/15/15 ratio. Thus, there are 14,229 records for training, 3,049 for validation, and 3,050 for testing.

Feature and target standardisation (zero mean, unit variance) is fitted on the training partition only and applied, unchanged, to the validation, test, OOD, and experimental data, precluding information leakage through the scalers. Since standardisation is a positive affine map,

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

with $\sigma > 0$, the sign of the partial derivative is preserved between the scaled and physical spaces. This property is load-bearing for the formulation of the constraint in 2.5.1. Row indices are tracked through the split so that the auxiliary capacity columns remain exactly aligned with each partition.

2.4. Baseline models

Four conventional learners are trained on identical splits to establish the accuracy attainable without physics guidance. XGBoost [31] was trained with 500 estimators, depth 6, learning rate 0.05, subsample and column-subsample of 0.8, and early stopping at 30 rounds against validation RMSE. LightGBM [32] was trained with matched settings. Random forest was trained with 500 trees and a minimum leaf size of 2. A standard multi-layer perceptron (64–64–32 hidden units, ReLU, Adam, L2 penalty 10^{-4}, early stopping) was trained on standardised inputs and targets, since plain MLPs are scale-sensitive. All baselines were evaluated on the validation, test, and OOD splits, in physical units. The MLP additionally serves as a loose architectural reference for the PGNN, while a strict architecture-matched control is provided within the ablation at section 2.6.

2.5. The physics-guided neural network

2.5.1. Architecture

The PGNN is a fully connected network with a shared trunk of three hidden layers (64–64–32 units) and two linear output heads. The primary head predicts the standardised fire resistance, FR. The auxiliary head predicts the standardised residual capacity. It exists so that the capacity-bound constraint (2.5.3) has a predicted quantity to constrain, which is anchored to the real final-capacity labels by a small auxiliary mean-squared-error term. Sharing the trunk lets the capacity task act as a mild multi-task regularisation without a separate network. The hidden activations use SiLU ($x \cdot \mathrm{sigmoid}(x)$) rather than ReLU [33], for a reason specific to physics guidance. The monotonicity penalties in 2.5.2 require differentiating the network output with respect to the input within the training loss, and then differentiating those penalties again with respect to the weights. ReLU's derivative is piecewise constant, so its second derivative vanishes almost everywhere and is exactly zero in the inactive region. This starves the penalty terms of the smooth gradient signal they need, whereas SiLU is C^∞ everywhere. Softplus was retained as a valid alternative.

2.5.2. Monotonicity constraints (gradient penalties)

Six sign relations between the inputs and fire resistance are established. Fire resistance does not decrease with increasing insulation thickness (t_{ins}), insulation depth (h_i), or glass-transition temperature (T_g), because more insulation or a more heat-tolerant adhesive cannot shorten the beam's survival time. Similarly, fire resistance cannot increase with increasing conductivity k_{ins}, applied load (L_d), or load ratio (LR), because more conductive insulation or a heavier load cannot increase resistance. For each training batch, the partial derivatives $\partial\widehat{\mathrm{FR}}/\partial x'_j$ are obtained in a single automatic-differentiation pass, and each wrong-signed derivative is hinge-penalised.

$$L_{\mathrm{mono}} = \sum_{j \in J^+} \mathrm{mean}\left(\mathrm{ReLU}\left(-\frac{\partial\widehat{\mathrm{FR}}}{\partial x'_j}\right)\right) + \sum_{j \in J^-} \mathrm{mean}\left(\mathrm{ReLU}\left(+\frac{\partial\widehat{\mathrm{FR}}}{\partial x'_j}\right)\right) \tag{2}$$

where J^+ and J^- index the features constrained to be increasing and decreasing in fire resistance, respectively. Penalties are computed directly on the standardised tensors, using the sign-preservation property established in 2.3. A sign constraint enforced in scaled space is therefore exactly a physical

constraint. The six channels are grouped into four weighted terms (insulation, conductivity, load, and T_g) so that their relative importance can be tuned independently.

2.5.3. Bound constraints

Two inequality constraints are enforced as hinges. First, predicted fire resistance must be non-negative. The scaled prediction is inverse-transformed to minutes, and the negative part is penalised. The penalty is divided by σ_y so that the term is dimensionless and of order one, like every other term. Second, predicted residual capacity at failure cannot exceed the beam's known ambient capacity. Both capacities are standardised with a single scaler fitted on the training final capacities (a shared positive affine map preserves the inequality), and $\mathrm{ReLU}\Big(\hat{C}_{\mathrm{final}} - C_{\mathrm{initial}}\Big)$ is penalised. This constraint is the auxiliary head's purpose.

2.5.4. Limit-state collocation constraint

A beam loaded at or beyond 100% of its ambient capacity fails essentially immediately upon exposure to fire. This implies that fire resistance must collapse toward zero as LR $\to$ 100%. However, the dataset never exceeds LR $\approx$ 73.5%, meaning that this law lives entirely outside the data. It is therefore enforced at synthetic collocation points, following standard physics-informed practice for imposing known laws in regions the data cannot cover. Each training batch is copied, inverse-transformed to physical units, and its LR column overwritten with values drawn uniformly from [100%, 150%]. These are re-standardised and passed through the network, and the prediction is penalised as a two-sided corridor about the limit state: an upper hinge on predictions exceeding a 5-minute tolerance, plus a lower hinge on predictions below zero (both normalised by σ_y),

$$L_{\mathrm{coll}} = \frac{1}{\sigma_y}\,\mathrm{mean}\Big(\,\mathrm{ReLU}\big(\widehat{\mathrm{FR}}_{\mathrm{coll}} - \delta\big) + \mathrm{ReLU}\big(-\widehat{\mathrm{FR}}_{\mathrm{coll}}\big)\Big), \qquad \delta = 5\ \mathrm{min}. \quad (3)$$

The lower hinge is not redundant with the global bound constraint of 2.5.3: that bound is evaluated on real training batches, whose predictions sit far from zero, so it exerts no restoring force in the collocation region. Without the lower wall, the load-ratio monotonicity penalty — which rewards a steeply decreasing response in LR — would be free to drive collocation-region predictions to large negative values while the one-sided tolerance hinge remained satisfied; the corridor closes this loophole and pins the limit-state

response inside $[0, \delta]$. The collocation LR values are resampled each batch during training. For any reported evaluation of this term, however, a fixed-seed sample is used so that the logged values are deterministic.

2.5.5. Asymmetric safety objective

The data-fit term replaces the symmetric MSE with a risk-weighted quadratic loss:

$$L_{\text{data}} = \text{mean}\Big(w\,(\widehat{\text{FR}} - \text{FR})^2\Big), \qquad w = \begin{cases} \alpha, & \text{if } \widehat{\text{FR}} > \text{FR}, \\ 1-\alpha, & \text{otherwise.} \end{cases} \tag{4}$$

with $\alpha = 0.8$, so an over-prediction of fire resistance — an unsafe error that could inform an inadequate design — is penalised four times as heavily as an under-prediction of equal magnitude. The loss is computed on the standardised tensors, since standardisation preserves order. The over-prediction indicator is identical in scaled and physical space, so the asymmetry's semantics are exact. By setting $\alpha = 0.5$, ordinary MSE is recovered, which is how the symmetric ablation cells are implemented.

2.5.6. Total loss and unit discipline

The full objective is

$$\begin{aligned} L_{\text{total}} = {} & L_{\text{data}} + \lambda_{\text{cap-fit}} L_{\text{cap-fit}} \\ & + r(t)\Big[\lambda_{\text{ins}} L_{\text{ins}} + \lambda_k L_k + \lambda_{\text{load}} L_{\text{load}} + \lambda_{T_g} L_{T_g} + \lambda_{\text{bound}} L_{\text{bound}} + \lambda_{\text{cap}} L_{\text{cap}} + \lambda_{\text{coll}} L_{\text{coll}}\Big]. \end{aligned} \tag{5}$$

where $r(t)$ is the warm-up ramp function described in Section 2.5.7, and L_{coll} is the limit-state collocation term of 2.5.4. The bracketed terms represent the four monotonicity constraints together with the two bound constraints and the collocation constraint. Two implementation principles were enforced throughout training. First, every loss term was expressed in standardised units of order one. Second, both the data-fitting and capacity-fitting objectives were evaluated on standardised tensors, while the two physical-unit hinge penalties were divided by σ_y. Without this normalisation, the individual loss terms differed by several orders of magnitude, making the weighting coefficients λ difficult to interpret. During preliminary experiments, an unnormalised capacity term dominated the combined gradient and overwhelmed

the primary prediction objective. After normalisation, the weighting coefficients became meaningful as relative importance factors. Unless otherwise stated, the default values were

$$\lambda_{\text{ins}} = \lambda_k = \lambda_{\text{load}} = \lambda_{\text{coll}} = 0.1, \quad \lambda_{T_g} = \lambda_{\text{bound}} = 0.05, \quad \lambda_{\text{cap}} = 0.2, \quad \lambda_{\text{cap-fit}} = 0.3.$$

The released implementation also provides an Optuna-based hyperparameter search [34] for automatic tuning of the λ coefficients.

2.5.7. Training protocol

The networks are trained with Adam [35] (learning rate 10^{-3}, weight decay 10^{-4} for parity with the baseline MLP's L2 penalty, and batch size 256) for up to 500 epochs, with a plateau-halving learning-rate schedule. Three protocol details are consequential:

1. **Physics warm-up:** Physics terms are linearly ramped from zero during the first 30 epochs, $r(t) = \min\left(1, \frac{t+1}{30}\right)$. Having all constraints active from initialisation results in a degenerate attractor — a nearly constant predictor that trivially satisfies all monotonicity and bound penalties but loses all predictive power — and a randomly initialised network sits close to this basin. The warm-up procedure allows the model first to learn the data manifold and then to be constrained.
2. **Model selection based only on the main task:** Early stopping (patience 40) and checkpointing are governed solely by the fire-resistance validation RMSE in minutes, not by the composite loss, which includes the auxiliary head and stochastically sampled collocation terms, since otherwise both mixing and sampling noise would enter the weight-selection procedure.
3. **Evaluation in physical units:** All results are reported after back-transformation into minutes.

2.6. Factorial ablation and multi-seed replication

The proposed approach differs from the traditional MLP architecture in two ways simultaneously — physics constraints and the asymmetric loss function — so a direct comparison would confound these two factors. These are separated in a 2×2 factorial ablation using shared architecture, optimiser, schedule and data as shown in Table 2.

Table 2: Ablation study configurations. The control is a standard MLP with symmetric loss and no physics constraints. The physics-only variant adds the physics terms to the control, while the asymmetric-only variant adds the asymmetric loss to the control. The full PGNN combines both physics and asymmetric loss.

Model	**Physics Terms**	**Data Loss**
Control	Off ($\lambda = 0$ for all terms)	Symmetric ($\alpha = 0.5$)
Physics-only	On	Symmetric
Asymmetric-only	Off	Asymmetric ($\alpha = 0.8$)
Full PGNN	On	Asymmetric

The control uses an auxiliary-fit weight of zero, thus making it a classic FR regressor without any additional constraint.

For each of these four settings, the network is trained using five different random seeds (42, 7, 123, 2024, and 31415) affecting weight initialisation, batch shuffling, and collocation sampling. The split is held constant across all seeds; this is intentional, so that differences between seeds isolate training randomness from split randomness. Split-level noise is acknowledged but cannot be quantified here (2.9). All metrics are reported as mean ± sample standard deviation (ddof = 1) across seeds. The main accuracy difference (physics vs. no physics) is also measured as a win count per seed pair: since paired variants share the same seed, this paired comparison is more powerful than comparing means alone. With $n = 5$, formal statistical hypothesis testing would be underpowered and is therefore intentionally avoided in favour of effect-size measurement; accordingly, no claim anywhere in this paper is stated as statistical significance, and overlapping results are described as overlapping at one standard deviation.

The factorial design fixes α at two levels, 0.5 and 0.8. To characterise the behaviour between and beyond them, the same training protocol is additionally run over an asymmetry sweep: six values of α (0.5, 0.6, 0.7, 0.8, 0.9, 0.95, corresponding to penalty ratios of 1.0:1 to 19.0:1) at the same five seeds, with the physics constraints active in every cell, giving 30 further training runs. Split, architecture, optimiser, schedule and warm-up are unchanged. Crucially, the asymmetry used at *evaluation* time is held fixed at $\alpha_{\text{eval}} = 0.8$ for every cell, so that safety loss compares models on one common risk preference instead of grading each model against its own training objective. Because the sweep contains the physics-only and full-PGNN cells of the factorial design as two of its six levels, it also serves as an independent re-execution check on

the tables reported below. Results are given in Section 3.6.

2.7. Physics violation rate and response analysis

2.7.1. Definition

Physical consistency is quantified using a metric identically applicable to every model class. Each of the six sign-constrained features is evaluated by perturbing that feature alone by +0.25 training standard deviations. A beam is counted as violating that constraint when the predicted fire resistance moves in the physically impossible direction by more than a 0.5-minute tolerance, to exclude numerical noise. The Physics Violation Rate (PVR) for a feature is the percentage of beams violating that constraint. For the headline ("any") constraint, PVR is the percentage of beams violating at least one constraint.

The finite perturbation is essential to fairness. Tree ensembles are almost everywhere piecewise-constant functions with zero gradient, so a gradient-based measure would be undefined for such models and incomparable across model classes. A finite perturbation asks each model the same question: "What happens if a designer adds a quarter-σ of insulation?" PVR is evaluated on both the test and OOD splits, using a seeded subsample of at most 3,000 beams per split for tractability, across all four ablation variants and XGBoost. Because the absolute rate depends on the perturbation size and tolerance, all models are measured under identical settings, so cross-model comparisons are invariant to this choice. The absolute rate is reported only after these settings are stated.

2.7.2. Response curves

The mechanism behind these rates is visualised using median-beam response curves. The feature-wise median beam from the test set is swept along one feature at a time, with all other features held fixed. Each model's predicted FR is traced, and two sweeps are reported. One sweeps insulation thickness over its training range; the other sweeps load ratio, extended to 150% — deliberately beyond the data maximum of 73.5% — with the extrapolation region marked. The latter exposes each model's extrapolation philosophy directly, including whether the PGNN's collocation-trained limit-state behaviour — in which FR collapses toward zero beyond LR = 100% — materialises in a region containing no training data.

2.8. Blind experimental validation protocol

The final phase evaluates every model against the 50 real furnace tests, under a designed protocol. A feature-envelope analysis revealed a substantial distributional gap between the simulation campaign and laboratory practice: the simulations hold f_y within 414–460 MPa and f_c within 27–40 MPa, whereas the test set spans 364–591 MPa and 23–52 MPa respectively. This gap is quantified as a result in 3.8.

1. **Grading, not gating:** Each real beam receives a continuous extrapolation-severity score, the worst per-feature exceedance of the training bounds, measured in training-σ units.

$$s_i = \max_j \left[\frac{\max(x_{\mathrm{lo},j} - x_{ij},\, x_{ij} - x_{\mathrm{hi},j},\, 0)}{\sigma_j} \right]. \tag{6}$$

   Note that s_i measures exceedance beyond the training *bounds*, normalised by σ_j; it is not a distance from the training mean. Beams are binned into three categories: mild ($s \leq 1$), moderate ($1 < s \leq 3$), and severe ($s > 3$). All metrics are reported per stratum, with no beam excluded, since discarding hard cases would constitute selective validation.
2. **Deployment guard:** Neural networks extrapolate approximately linearly in standardised space, so a real input several σ outside the training range can produce unbounded outputs. Tree ensembles are immune to this only because their leaf structure implicitly clamps outputs. An explicit input-clipping guard, in which each feature is clamped to its training range before prediction, is therefore applied to all models, levelling the comparison; metrics are reported both raw and guarded. The guard's known distortion — that a clamped feature causes the model to effectively predict a modified beam — is acknowledged. For fire resistance, it acts predominantly in the conservative direction.
3. **Directional metric suite:** We report RMSE, MAE, and R^2 — flagged as degenerate for the $n = 3$ mild stratum — along with mean signed error and bias-removed RMSE ($\sqrt{\mathrm{RMSE}^2 - \mathrm{bias}^2}$), for every model and stratum. The count and mean magnitude of over-predictions, the worst-case over-prediction, and the asymmetric safety loss at $\alpha = 0.8$ are applied uniformly across all models, so that the metric suite is risk-weighted and applied consistently, making cross-model comparison fair.

The blindness of this evaluation is structural: the split, scalers, and checkpoints are all fixed in advance. All thresholds derive exclusively from the simulation corpus, and the experimental dataset enters the pipeline only at prediction time.

2.9. Scope and threats to validity

These four methodological limitations are stated for completeness. The seed replication ($n = 5$) is sufficient for effect-size estimation and paired comparisons but not for formal hypothesis testing. The fixed data partition leaves split-level variance unquantified. The absolute value of PVR is settings-dependent, although cross-model comparisons are not. The asymmetry level $\alpha = 0.8$ encodes a single risk preference. The framework accepts any value of α, and the conservatism–accuracy frontier is mapped over six levels in Section 3.6; that sweep is nevertheless run with the physics constraints active throughout, so α and the constraint set are separated only by combining it with the factorial ablation. Codes and respective results are released for exact reproduction[36].

3. Results and Discussion

3.1. Baseline performance and the extrapolation problem

Table 3: Performance of conventional baseline models on the validation, test, and out-of-distribution (OOD) extrapolation splits. RMSE and MAE in minutes.

	Validation			Test			OOD		
Model	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2
XGBoost	20.39	11.75	0.925	22.19	12.51	0.912	38.13	25.06	0.455
LightGBM	21.81	13.09	0.914	23.98	13.98	0.898	38.09	24.90	0.456
Random forest	21.87	11.86	0.913	23.82	12.61	0.899	50.55	30.68	0.042
MLP	23.61	14.76	0.899	25.12	15.25	0.888	44.06	28.25	0.272

Table 3 shows the metrics of four baseline models (XGBoost, LightGBM, random forest and MLP) on validation, test and OOD splits. The four conventional learners establish the accuracy attainable without physics guidance. On the in-distribution test split, XGBoost performed best (RMSE = 22.2 min, MAE = 12.5 min, $R^2 = 0.912$), closely followed by random forest (RMSE = 23.8 min, $R^2 = 0.898$), with the standard MLP slightly behind

(RMSE = 25.1 min, $R^2 = 0.888$). All four models interpolated the simulated fire-resistance response well. For designers querying configurations similar to those in the simulation campaign, any of these models would appear adequate, and the gradient-boosted ensembles would appear excellent. The out-of-distribution (OOD) split dismantles this impression. On beams held out for having a load ratio in the top 5% of the corpus — a mild extrapolation by engineering standards, since these are physically ordinary designs — every baseline degraded severely. XGBoost and LightGBM fell to $R^2 \approx 0.46$ (RMSE $\approx$ 38 min), the MLP to $R^2 = 0.27$ (RMSE = 44.1 min), and random forest collapsed almost entirely to $R^2 = 0.04$ (RMSE = 50.6 min). The failure of random forest is instructive: bagged trees cannot predict beyond the range of their training targets, so the model saturates [37]. Boosted ensembles degrade more gracefully but still lose half their explanatory power. This quantitative expression of the problem motivates the present work, because high in-distribution accuracy in a black-box surrogate says nothing about its behaviour in regions where the design tool is most likely to be queried and least likely to be checked.

3.2. The factorial ablation: physics guidance is accuracy neutral

Table 4: Factorial ablation on the in-distribution test split (mean ± std over five seeds). Bias is the mean signed error; Over is the over-prediction rate; MaxOver is the worst-case over-prediction; L_s is the asymmetric safety loss ($\alpha = 0.8$) evaluated identically for all variants. RMSE, bias, MaxOver in minutes; L_s in min^2.

Variant	RMSE	R^2	Bias	Over (%)	MaxOver	L_s
Control (no physics, sym.)	27.05 ± 1.41	0.870 ± 0.014	$+0.13 \pm 0.32$	49.1 ± 0.7	262.6 ± 7.2	433.2 ± 44.1
Physics only (sym.)	26.28 ± 0.55	0.877 ± 0.005	$+0.27 \pm 0.29$	48.5 ± 0.6	268.4 ± 6.8	414.8 ± 17.4
Asymmetric only	34.10 ± 0.22	0.793 ± 0.003	-11.61 ± 0.45	30.0 ± 0.8	269.0 ± 11.1	460.0 ± 8.0
PGNN (physics + asym.)	33.81 ± 0.48	0.797 ± 0.006	-11.83 ± 0.69	30.2 ± 0.9	262.5 ± 10.2	445.8 ± 11.2

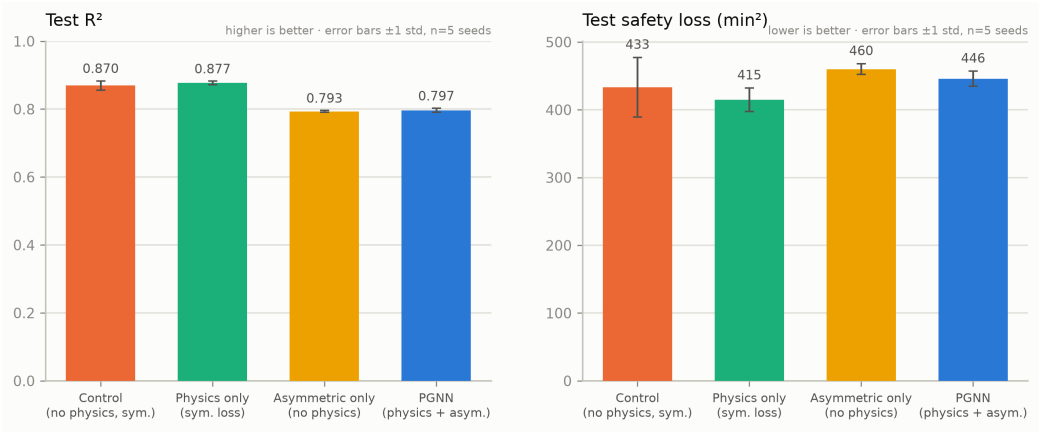


Figure 2: Test-split performance across the four cells of the 2×2 factorial ablation (Table 4): Control (no physics, symmetric loss), Physics-only (symmetric loss), Asymmetric-only (no physics), and full PGNN (physics + asymmetric loss). Bars show means over five random seeds with ±1 sample standard deviation. The left panel reports R^2 (higher is better), while the right reports asymmetric safety loss L_s ($\alpha = 0.8$, min^2; lower is better). Physics constraints have only a marginal effect on R^2, whereas the asymmetric objective produces the observed accuracy reduction. Aggregate safety loss shows substantial overlap across variants, motivating the directional safety metrics reported in Table 4.

The 2 × 2 ablation — {physics off/on} × {symmetric/asymmetric loss}, five seeds per cell — as represented in Table 4, separates the two mechanisms that the full PGNN combines. Figure 2 shows a bar chart for test R^2 and safety loss for ablation variants within ±1 std error bars. The first finding concerns the cost of physics guidance, and the answer is none. The physics-constrained symmetric variant achieved test $R^2 = 0.8771 \pm 0.0051$, against 0.8696 ± 0.0137 for its unconstrained twin, with RMSE of 26.3 ± 0.6 versus 27.1 ± 1.4 min. The mean difference ($\Delta R^2 = +0.008$) lies within one standard deviation of the control's spread, and in a paired-by-seed comparison — a more sensitive test, since paired variants share initialisation — the physics variant exceeded the control in two of five seeds, with the three losses each smaller than 0.003 and the two wins as large as 0.018 and 0.024. This implies a deliberately bounded claim: the constraints neither systematically improve nor degrade interpolation accuracy. We do not claim improvement, because a split seed-wise record does not establish it, but we do claim cost-freeness, because the positive mean difference and the overlapping distributions exclude any meaningful penalty. It is also notable that the physics variant is markedly more stable across seeds (± 0.005 against ± 0.014 in test R^2), a

first appearance of the dispersion effect analysed in Section 3.5. Neutrality is not expected under the common presumption that physics penalties compete with data fit for model capacity. A plausible mechanism is that monotonicity constraints act as a physically aligned regulariser [29]: they exclude hypotheses inconsistent with the true function, so the restriction removes variance without introducing bias. This finding plays a structural role in the rest of the analysis, because physics accounts for none of the remaining accuracy difference. The entire R^2 gap between control (0.870) and full PGNN (0.797) is therefore attributable entirely to the asymmetric safety objective. This gap represents a purchase, and the next subsection itemises what it bought.

3.3. The asymmetric objective: pricing conservatism explicitly

The asymmetric variants tell a consistent story across both cells of the ablation. Training under 4:1 risk weighting shifts the error distribution toward the safe side. The full PGNN's test bias is -11.8 ± 0.7 min (control: $+0.1 \pm 0.3$), and its over-prediction rate falls from the $49.1 \pm 0.7\%$ expected of a symmetrically-trained estimator to $30.2 \pm 0.9\%$.

Decomposing RMSE^2 into bias^2 plus variance shows that roughly one-third of the apparent accuracy penalty is the deliberate shift itself. Removing the bias leaves a centred RMSE of ≈ 31.7 min against the control's 27.1 min, so 4.6 min of the 6.8 min nominal gap survives debiasing. In R^2 terms, a debiased PGNN would score ≈ 0.821 against the control's 0.870, so about 0.024 of the 0.073 R^2 gap is pure bias and the remainder is increased dispersion.

In design terms, the PGNN embeds a soft safety allowance of approximately 12 minutes directly into its predictions. A designer using the symmetric control would need to impose an equivalent margin externally and apply it uniformly, whereas the learned asymmetry concentrates conservatism where the model's own uncertainty is largest. Two evaluation consequences follow. First, R^2 systematically misrepresents a deliberately conservative estimator, because it rewards unbiased mean prediction — the very property the training objective was designed to avoid. Comparisons between symmetric and asymmetric models must therefore be accompanied by directional metrics. Second, on the metric that encodes the domain's actual risk preferences — the asymmetric safety loss, evaluated identically for all variants — the aggregate ordering turns out to be largely uninformative within the training distribution. On the test split, physics-only attains the numerically lowest value (414.8 ± 17.4 min^2), followed by control (433.2 ± 44.1) and PGNN (445.8 ± 11.2), with asymmetric-only highest (460.0 ± 8.0); each adjacent

pair overlaps at one standard deviation. The validation split gives the same picture and the same ordering: physics-only 314.2 ± 10.0 min^2, control 333.4 ± 40.8, PGNN 348.1 ± 11.6, and asymmetric-only 355.0 ± 7.5. On neither split does the aggregate L_s separate the asymmetric variants from the symmetric ones in the direction the training objective might suggest. The more informative comparison is directional: despite its nominal R^2 disadvantage, the full PGNN reaches a comparable safety loss to the control while cutting the over-prediction (unsafe-error) rate by nearly 40% (30.2% vs. 49.1%), a difference the aggregate L_s value does not by itself reveal. The interaction between the two mechanisms is also resolved by the ablation, under the asymmetric objective. Adding physics improves or matches every reported metric relative to the asymmetric-only variant: test R^2 (0.797 vs 0.793), test safety loss (445.8 vs 460.0), OOD R^2 (0.219 vs 0.182), and OOD worst-case over-prediction (80.8 vs 87.8 min), with the test over-prediction rate the one metric on which the two are level (30.2% vs 30.0%). Each individual difference is modest, but their consistency across metrics and splits indicates complementarity. The asymmetric loss shapes the error distribution, while the physics constraints shape the learned function.

3.4. Out-of-distribution behaviour: the trade made explicit

Table 5: Factorial ablation on the OOD extrapolation split (mean ± std over five seeds). Columns as in Table 4. Note the roughly sevenfold reduction in the seed dispersion of L_s for the full PGNN (constraint-induced reliability).

Variant	RMSE	R^2	Bias	Over (%)	MaxOver	L_s
Control (no physics, sym.)	39.37 ± 1.73	0.418 ± 0.051	-12.06 ± 2.37	40.2 ± 3.2	143.8 ± 36.9	490.2 ± 106.8
Physics only (sym.)	39.54 ± 1.80	0.413 ± 0.053	-13.83 ± 1.30	36.4 ± 1.8	154.4 ± 29.1	465.6 ± 41.4
Asymmetric only	46.73 ± 0.28	0.182 ± 0.010	-25.24 ± 0.47	16.8 ± 1.6	87.8 ± 8.5	484.3 ± 8.8
PGNN (physics + asym.)	45.65 ± 1.00	0.219 ± 0.034	-25.10 ± 0.73	$\mathbf{15.5 \pm 1.0}$	$\mathbf{80.8 \pm 15.6}$	$\mathbf{458.4 \pm 15.0}$

On the OOD split, as shown in Table 5, the full PGNN's R^2 (0.219 ± 0.034) lies well below the control's (0.418 ± 0.051), and we state plainly that this is systematic rather than noise. The relevant question is the direction of each model's failure. The PGNN fails conservatively: its OOD bias is -25.1 ± 0.7 min, its over-prediction rate is 15.5 ± 1.0% against the control's 40.2 ± 3.2%, and its worst-case over-prediction is 80.8 ± 15.6 min against 143.8 ± 36.9 min — a 44% reduction in the single most safety-critical statistic in this study. On risk-weighted safety loss, the PGNN is the best variant on this split (458.4 ± 15.0 vs 490.2 ± 106.8 min^2). For surrogate screening of

candidate designs, an extrapolation regime that errs systematically toward shorter predicted fire resistance is the correct failure mode. It may waste insulation, but it cannot silently certify an inadequate design. The OOD result is therefore not a weakness to be excused, but a quantified, intended operating characteristic. Approximately 0.20 of OOD R^2 was exchanged for cutting the unsafe-error rate by over 60% and reducing the worst unsafe error by more than two-fifths.

3.5. Constraint-induced reliability: the variance collapse

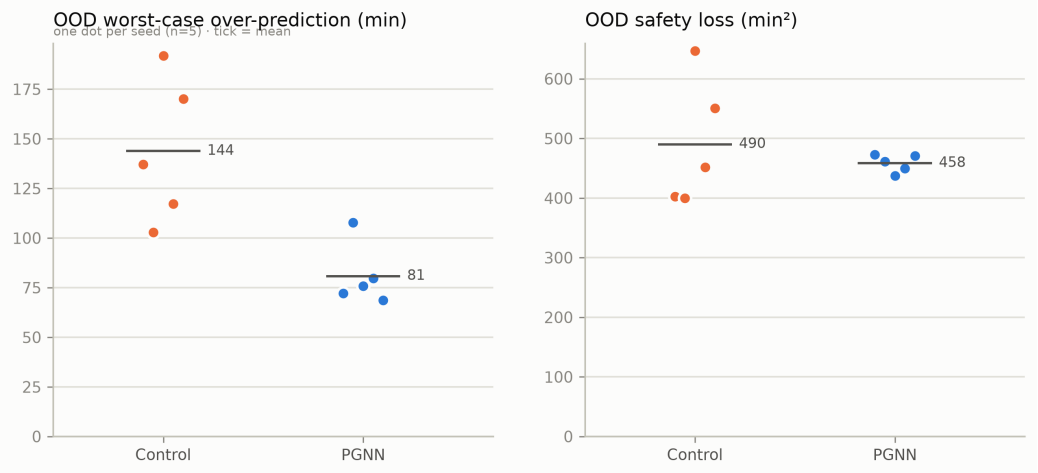


Figure 3: Strip plots of OOD worst-case over-prediction and OOD safety loss across five seeds, control vs. PGNN. Each point represents one random seed; the horizontal tick marks the mean of the five seeds for that variant.

An unanticipated finding concerns the dispersion of safety metrics across random initialisations, not their mean values. Across five seeds, the unconstrained control's OOD safety loss varied by ± 107 min^2 and its worst-case over-prediction by ± 37 min—substantial variation for models differing only in random seed. The PGNN reduced these to ± 15.0 min^2 and ± 15.6 min, a roughly sevenfold reduction in safety-loss dispersion and better than twofold for worst-case over-prediction. Similar reductions appear in OOD bias (± 0.7 vs ± 2.4 min) and over-prediction rate (± 1.0 vs ± 3.2 percentage points).

Per-seed OOD MaxOver values for the PGNN are 72.0, 107.7, 75.9, 79.8, and 68.6 min, giving an honest bound of roughly 110 min (four of five seeds stay below 80 min). The unconstrained control's per-seed worst cases reach 192 min.

We call this reduction in seed-dependent dispersion *constraint-induced reliability*. Here it is quantified through the variability of OOD safety loss, bias, over-prediction rate, and worst-case over-prediction across seeds. Seed-dependent variability is seldom examined in surrogate modelling, but the results suggest it should be part of evaluating safety-critical models, especially when their use involves extrapolation beyond the labelled data domain.

The comparison above conflates two mechanisms. Separating them across the middle ablation cells, the symmetric physics-only variant reduces OOD safety-loss dispersion from ± 106.8 to ± 41.4 min^2, while the asymmetric-only variant reduces it to ± 8.8 min^2 with no physics terms at all. The asymmetry sweep in Section 3.6 confirms this: holding constraints active and raising α from 0.5 to 0.8 alone collapses the dispersion from ± 39.9 to ± 14.6 min^2. Constraint-induced reliability, as measured here, is therefore driven predominantly by the asymmetric objective, with physics terms contributing a smaller, separable share. The term should not be read as a property of the monotonicity and collocation constraints in isolation.

3.6. Sensitivity to the asymmetry parameter

Table 6: Asymmetry sweep. Six values of α, five seeds each (mean $\pm$ std), with the physics constraints active in every cell and the evaluation asymmetry held fixed at $\alpha_{\text{eval}} = 0.8$ so that L_s is a common yardstick rather than each model's own training objective. The $\alpha = 0.5$ and $\alpha = 0.8$ rows are the physics-only and full-PGNN cells of Tables 4 and 5 respectively, and reproduce them to within 1.2% on every reported metric. Bias, MaxOver and L_s are in minutes, minutes and min^2.

		Test (in-distribution)			OOD extrapolation				
α	Ratio	R^2	Over (%)	L_s	R^2	Bias	Over (%)	MaxOver	L_s
0.50	1.0:1	0.877 ± 0.005	48.5 ± 0.7	415.5 ± 17.8	0.413 ± 0.053	-13.9 ± 1.5	36.1 ± 2.0	153.4 ± 27.4	464.0 ± 39.9
0.60	1.5:1	0.862 ± 0.007	44.0 ± 0.3	419.8 ± 19.0	0.363 ± 0.061	-18.1 ± 1.8	29.0 ± 2.2	124.1 ± 13.1	437.4 ± 26.3
0.70	2.3:1	0.832 ± 0.010	39.0 ± 1.3	452.4 ± 22.4	0.338 ± 0.016	-21.1 ± 0.4	20.6 ± 1.4	99.2 ± 13.7	$\mathbf{417.1 \pm 10.4}$
0.80	**4.0:1**	0.797 ± 0.006	30.2 ± 0.8	445.8 ± 11.1	0.218 ± 0.034	-25.1 ± 0.7	15.4 ± 1.0	80.8 ± 15.9	458.8 ± 14.6
0.90	9.0:1	0.692 ± 0.013	19.7 ± 0.6	500.0 ± 15.7	0.093 ± 0.044	-29.8 ± 1.2	10.4 ± 0.8	79.4 ± 15.8	514.5 ± 19.6
0.95	19.0:1	0.529 ± 0.023	13.8 ± 0.5	631.8 ± 23.1	-0.079 ± 0.047	-34.9 ± 1.0	7.1 ± 0.4	66.0 ± 7.9	595.9 ± 26.3

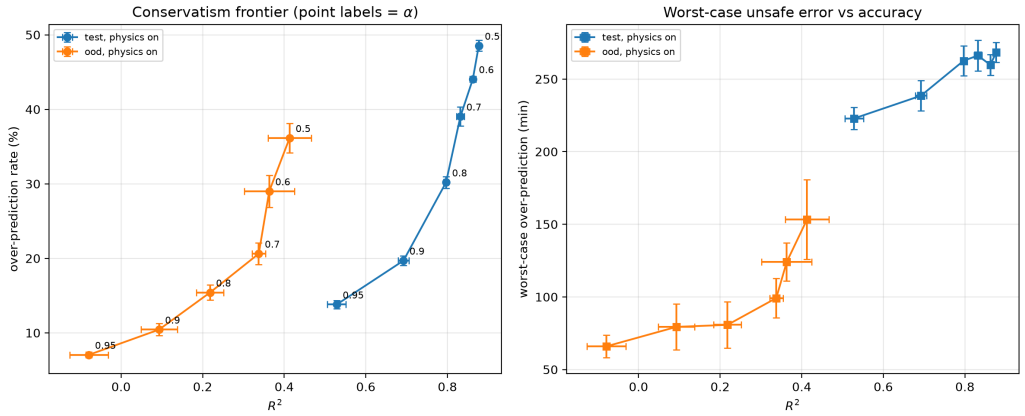


Figure 4: Conservatism–accuracy frontier traced by the asymmetry parameter: over-prediction rate vs. R^2 (a) and worst-case over-prediction vs. R^2 (b), each shown for both the in-distribution test split and the OOD extrapolation split. Each point is the mean over five seeds, with ± 1 std error bars in both coordinates; the R^2 axis is reversed so that accuracy decreases to the right. Points are labelled by α.

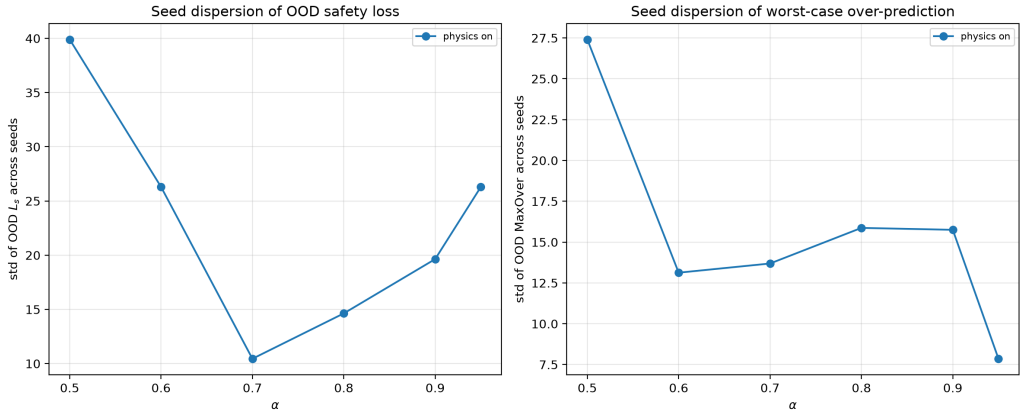


Figure 5: Out-of-distribution risk-weighted loss (a) and seed dispersion of the two safety statistics (b) as functions of α, with the physics constraints held active throughout. The dispersion of OOD safety loss is minimised at $\alpha = 0.7$ (± 10.4 min^2), while MaxOver dispersion falls rapidly at $\alpha = 0.95$.

The value $\alpha = 0.8$ used throughout the ablation encodes one particular risk preference, and a single point on a trade-off curve cannot justify itself. The curve is therefore traced directly. The network was retrained at six

asymmetry levels spanning symmetric loss to a 19:1 penalty ratio, five seeds per level, with the physics constraints active in every cell and the split, architecture, optimiser and schedule identical to Section 2.6. Because two of the six levels coincide with cells of the published ablation, the sweep doubles as a reproducibility check: the $\alpha = 0.5$ and $\alpha = 0.8$ rows of Table 6 recover the physics-only and full-PGNN rows of Tables 4 and 5 to within 1.2% on every reported metric across all three splits (the sole exception being the near-zero test bias of the physics-only cell, which agrees to within 0.03 min), confirming that the pipeline reported here is deterministic under re-execution.

The first result is that conservatism is not free, and its price is now measured rather than asserted. Accuracy falls monotonically in α on both splits: test R^2 declines from 0.877 to 0.529 and test RMSE nearly doubles, from 26.3 to 51.5 min, across the range examined. What matters for design is the exchange rate. Taking the symmetric model as the reference, the number of percentage points of unsafe prediction eliminated per 0.01 of test R^2 surrendered is 3.00 at $\alpha = 0.6$, 2.11 at 0.7, 2.29 at 0.8, 1.56 at 0.9 and 1.00 at 0.95. The rate remains above two up to $\alpha = 0.8$ and then falls away, so the frontier in Figure 4 has a knee at approximately that value. Beyond it, each further increment of conservatism costs roughly twice as much accuracy as the increments before it, and at $\alpha = 0.95$ the model ceases to have any OOD predictive skill at all ($R^2 = -0.079 \pm 0.047$, worse than predicting the training mean). The adopted value sits at the last point on the efficient portion of the curve, which is the justification the original choice lacked.

The second result concerns where the asymmetric objective actually pays. Within the training distribution it does not: test safety loss moves only mildly and non-monotonically across $\alpha = 0.5$ through 0.8 (415.5 ± 17.8, 419.8 ± 19.0, 452.4 ± 22.4 and 445.8 ± 11.1 min^2), a change small compared with the collapse of the unsafe statistics. This is consistent with, and extends, the finding of Section 3.3 that the aggregate L_s ordering is largely uninformative in-distribution: the asymmetric term does not reduce risk-weighted error on data the model has seen, it only relocates that error to the safe side. The entire benefit appears out of distribution, where OOD worst-case over-prediction falls monotonically from 153.4 ± 27.4 min at $\alpha = 0.5$ to 66.0 ± 7.9 min at $\alpha = 0.95$, and OOD L_s traces a genuine minimum (Figure 5a). The asymmetric loss is therefore best understood as an extrapolation-control device rather than an in-distribution accuracy device.

That minimum requires an honest statement, because it does not fall at the adopted value. On mean OOD safety loss the best cell is $\alpha = 0.7$ ($417.1 \pm$

10.4 min^2) rather than $\alpha = 0.8$ (458.8 ± 14.6), a gap of 41.7 min^2 against a combined standard error of 8.0, and $\alpha = 0.7$ also retains substantially more OOD accuracy ($R^2 = 0.338$ versus 0.218). Judged on the mean alone, $\alpha = 0.7$ dominates the configuration adopted here. Three considerations nonetheless favour $\alpha = 0.8$ for the intended use.

The first is the worst case. Mean OOD MaxOver is 80.8 ± 15.9 min at $\alpha = 0.8$ against 99.2 ± 13.7 min at $\alpha = 0.7$, a 19% reduction. Because a deployed model is one draw from the seed ensemble rather than its average, the envelope over seeds also matters: the largest OOD over-prediction observed at any seed is 108.2 min at $\alpha = 0.8$ and 119.2 min at $\alpha = 0.7$. The envelope difference is modest (10%), so the worst-case argument rests mainly on the mean; the bound of approximately 110 min reported in Section 3.5 holds at $\alpha = 0.8$ and is exceeded at $\alpha = 0.7$, though not by a margin that would by itself settle the choice.

The second is that the reversal is confined to one metric on one split. On both in-distribution splits the safety-loss ordering runs the other way (348.1 against 350.9 min^2 on validation, 445.8 against 452.4 on test), and on the over-prediction rate — the primary safety statistic used throughout this study — $\alpha = 0.8$ leads on all three splits (30.2 vs 38.9% on validation, 30.2 vs 39.0% on test, 15.4 vs 20.6% on OOD), as does MaxOver. The higher OOD L_s at $\alpha = 0.8$ is moreover incurred on the safe side of the ledger, since L_s retains a weight of $1 - \alpha$ on under-prediction and the OOD bias deepens from -21.1 to -25.1 min while every unsafe statistic improves.

The third is selection integrity. The ratio 4:1 was fixed in advance of this sweep, which is a post-hoc sensitivity analysis rather than a tuning procedure, and the OOD split is the held-out extrapolation probe against which the model is judged. Re-selecting α at its argmin on that split would convert the probe into a model-selection set and forfeit the property that makes it informative. Since a surrogate used for design screening is qualified on its tail behaviour and on the reproducibility of that behaviour across independently trained instances, and since in deployment the extrapolation region affords no labels against which a mean could be checked, $\alpha = 0.8$ is retained. A user whose loss function is genuinely the mean rather than the tail should prefer $\alpha = 0.7$, and the framework supports that choice without modification.

The third result revises the attribution made in Section 3.5. Holding the physics constraints active and varying only α, the seed dispersion of OOD safety loss falls from ± 39.9 min^2 at $\alpha = 0.5$ to ± 14.6 min^2 at $\alpha = 0.8$, and

the dispersion of OOD worst-case over-prediction falls from ± 27.4 to ± 15.9 min over the same interval — a factor of roughly 2.7 in the first case, produced by the loss asymmetry alone (Figure 5b). Read together with the factorial ablation, in which the symmetric physics-only variant reduces OOD L_s dispersion from ± 106.8 to ± 41.4 min^2 while the asymmetric-only variant reduces it to ± 8.8, the decomposition is clear: the physics constraints contribute roughly a factor of 2.6 and the asymmetric objective roughly a factor of 12. The variance collapse named constraint-induced reliability in Section 3.5 is therefore driven predominantly by the asymmetric objective, with the physics terms contributing a smaller, independent share. The dispersion is also non-monotone in α: within the sweep it is minimised at $\alpha = 0.7$ (± 10.4 min^2), sits only slightly higher at $\alpha = 0.8$ (± 14.6), and rises again toward both ends of the range, so reproducibility is a shared property of the intermediate asymmetry levels rather than a further argument that separates $\alpha = 0.8$ from $\alpha = 0.7$.

Two limitations of the sweep are stated. The physics constraints are active in every cell, so α and the constraint set cannot be fully separated within the sweep itself; the decomposition above depends on combining it with the 2×2 ablation, where the physics-off arm is available. And the OOD split is a load-ratio tail (Section 2.3), the same region in which the limit-state collocation term is designed to drive predicted fire resistance toward zero. Increasing α and enforcing the collocation constraint push predictions in the same direction there, so part of the OOD improvement attributed to α reflects the two terms acting in concert rather than the asymmetry acting alone. Neither limitation affects the in-distribution frontier or the exchange rates reported above.

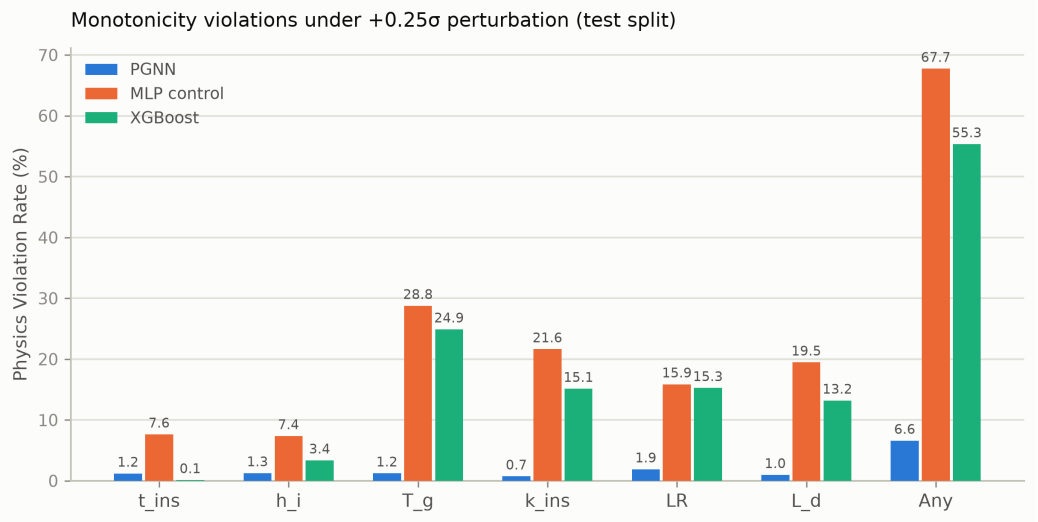


Figure 6: Physics Violation Rate (PVR) by constrained feature for PGNN, the unconstrained MLP control, and XGBoost on the test split. Each feature is perturbed individually by $+0.25\ \sigma$ while holding other inputs fixed, with a 0.5-minute tolerance used to identify violations (2.7.1). The "Any" group denotes beams violating at least one of the six constraints. PGNN reduces violations across all constrained features to below 2%, with an overall PVR of 6.6%, compared with 67.7% for the MLP control and 55.3% for XGBoost. Results for the physics-only and asymmetric-only ablations are reported in Table 7.

The Physics Violation Rate finding reframes the entire accuracy comparison. It is expressed in Table 7 and three representative models are shown visually in Figure 6. Under a $+0.25\sigma$ perturbation of each sign-constrained feature, with a 0.5-min tolerance excluding numerical noise, the unconstrained control produced a physically impossible response: fire resistance was predicted to fall when insulation was added, or to rise when load increased, for at least one variable on 67.7% of test beams. XGBoost, the accuracy leader in 3.1, violated at least one constraint for 55.3% of beams. A feature-level breakdown locates the damage: the control's largest violation channels are T_g (28.8%), k_{ins} (21.6%), and the two load variables (15.9% and 19.5%). XGBoost mirrors this pattern (T_g 24.9%, k_{ins} 15.1%, and LR 15.3%). These are not peripheral features; they are the variables a designer would manipulate directly when sizing an insulation scheme. A surrogate is queried for design purposes precisely through such perturbations. A 55–68% violation rate

Table 7: Physics Violation Rate (%): share of beams whose predicted FR responds in the physically impossible direction to a $+0.25\sigma$ perturbation of each constrained feature (tolerance 0.5 min). "Any" = at least one constraint violated.

Split	Model	t_{ins}	h_i	T_g	k_{ins}	LR	L_d	Any
Test	MLP control	7.6	7.4	28.8	21.6	15.9	19.5	67.7
	XGBoost	0.1	3.4	24.9	15.1	15.3	13.2	55.3
	Asymmetric only	1.6	1.3	17.5	9.1	11.3	7.2	36.7
	Physics only	4.0	3.5	3.8	1.3	3.7	2.9	16.4
	PGNN	1.2	1.3	1.2	0.7	1.9	1.0	**6.6**
OOD	MLP control	0.6	6.6	7.2	22.2	27.4	15.7	56.5
	XGBoost	0.0	5.2	16.3	17.9	0.0	12.1	40.5
	Asymmetric only	0.0	0.2	8.7	1.8	19.0	7.2	27.2
	Physics only	0.6	2.8	0.7	0.6	3.0	1.5	8.6
	PGNN	0.0	0.0	0.2	0.0	0.0	0.0	**0.2**

therefore means that the model gives a wrong-signed answer to the tool's primary use case for most beams, regardless of its R^2. Use of the PGNN suppresses the any-constraint rate to around 6.6% on the test split, with no individual feature exceeding 1.9%. The ablation decomposes this effect more cleanly. Physics constraints alone reach 16.4%, while the asymmetric objective alone reaches 36.7% — a genuine secondary effect, plausibly because depressing predictions flattens spurious positive gradients. The combination thus compounds to 6.6%. On the OOD split, the contrast sharpens to its extreme. The control remains at 56.5% and XGBoost at around 40.5%, while the PGNN falls to 0.2%, achieving exactly 0.0% on five of the six constraints. The unconstrained models are most unphysical precisely where they are least testable, whereas the PGNN's constraints continue to bind there, being properties of the function rather than of the data.

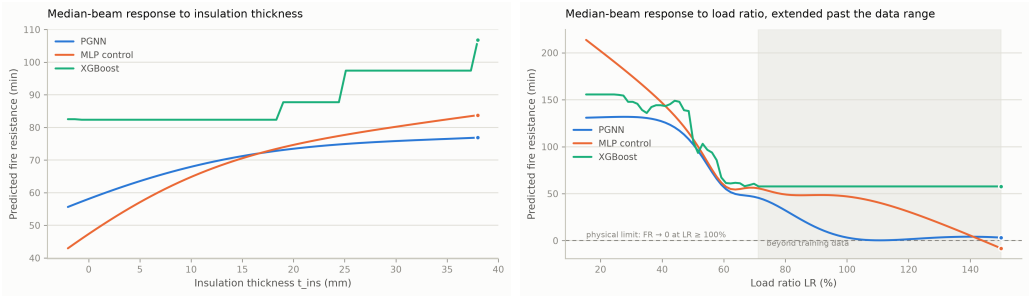


Figure 7: Median-beam response curves: predicted FR against t_{ins} (left) and against LR extended to 150% (right), with the extrapolation region shaded and the physical limit FR $\rightarrow$ 0 marked.

Response curves make the mechanism more visible as shown in Figure 7. The PGNN's fire-resistance response rises monotonically with insulation thickness and falls monotonically with load ratio, bending toward the collocation-trained collapse (FR settling within the 0–5 min corridor) as LR crosses 100% — a region containing no training data whatsoever, where the model's behaviour is supplied entirely by the encoded limit-state law. The control and XGBoost both exhibit non-monotonic excursions, within and beyond the data range alike.

Table 8: Blind validation against 50 real furnace tests. Raw predictions and predictions under the input-clipping deployment guard, with guarded results additionally stratified by extrapolation severity (moderate: 1–3σ, n=7; severe: >3σ, n=40; the mild stratum, n=3, is omitted as statistically degenerate). #Over = number of over-predicted (unsafe) beams. RMSE, bias, MaxOver in minutes; L_s in min^2. Raw and guarded results coincide for XGBoost, whose leaf structure clamps out-of-range inputs implicitly.

Subset	Model	RMSE	R^2	Bias	#Over	MaxOver	L_s
All, raw (n=50)	MLP control	254.7	−20.70	+148.1	38	618.2	50 570.8
	Asymmetric only	205.3	−13.09	+79.0	33	680.2	32 462.7
	PGNN	122.3	−4.00	+46.1	29	325.0	11 087.0
	Physics only	286.1	−26.37	+151.1	39	773.4	64 664.8
	XGBoost	50.2	+0.16	−16.4	23	60.8	701.3
All, guarded (n=50)	MLP control	70.3	−0.65	−22.0	18	210.2	1 940.7
	Asymmetric only	58.3	−0.14	−27.7	15	104.6	962.1
	PGNN	56.7	−0.08	−36.9	**13**	**31.7**	692.3
	Physics only	61.5	−0.27	−20.5	19	145.4	1 283.2
	XGBoost	50.2	+0.16	−16.4	23	60.8	701.3
Moderate, guarded (n=7)	MLP control	59.5	−0.39	+19.6	5	117.0	2 229.3
	Asymmetric only	64.6	−0.64	+2.5	3	104.6	2 220.9
	PGNN	39.6	**+0.38**	−13.2	3	31.7	503.9
	Physics only	40.9	**+0.34**	−6.3	3	55.0	664.5
	XGBoost	75.8	−1.26	−68.8	0	0.0	1 147.8
Severe, guarded (n=40)	MLP control	70.4	−1.77	−27.0	12	210.2	1 882.2
	Asymmetric only	54.9	−0.69	−30.7	11	38.1	701.4
	PGNN	57.1	−0.82	−39.6	9	24.2	671.4
	Physics only	62.8	−1.21	−19.9	15	145.4	1 387.3
	XGBoost	42.5	−0.01	−3.8	23	60.8	608.1

3.8. Blind experimental validation: a graded sim-to-real stress test

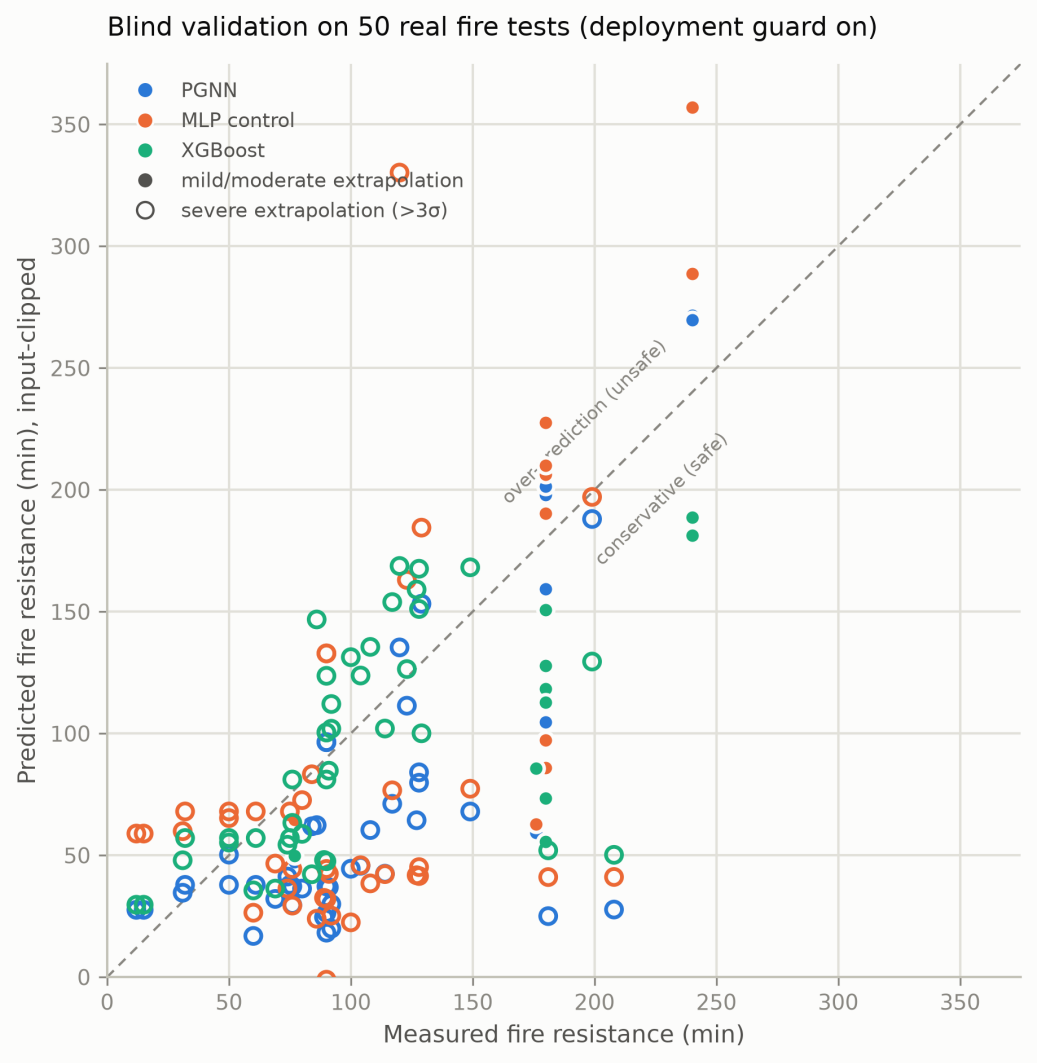


Figure 8: Parity plot of predicted vs. measured fire resistance (FR) under the input-clipping deployment guard, for three representative models (PGNN, MLP control, XGBoost). Markers are coded by extrapolation severity: filled markers denote beams within the mild or moderate strata ($s \leq 3\sigma$, n = 10 combined), and open circles denote the severe stratum ($s > 3\sigma$, n = 40); see Table 8 for metrics computed separately on the moderate and severe strata. The dashed diagonal is the line of perfect prediction; points above it are unsafe over-predictions, points below it are conservative.

The fifty real furnace tests, blind to every phase of model development, occupy a different region of feature space from the simulation campaign. After unit harmonisation of the source records, none of the 50 beams lay fully inside the training envelope. This gap is driven predominantly by the material properties: simulations held f_y within 414–460 MPa and f_c within 27–40 MPa, whereas the laboratory literature spans 364–591 MPa and 23–52 MPa, respectively, placing f_y out of range for 43 of the 50 beams. Graded by worst per-feature exceedance in training-σ units, 3 beams are mild extrapolations, 7 moderate, and 40 severe ($> 3\sigma$). Validation is therefore, unavoidably, a stress test of extrapolation behaviour; we report it as such, stratified, with no beam excluded. Table 8 shows the experimental validation of 50 real fire tests with raw and input-clipped metrics stratified by extrapolation severity, while Figure 8 shows a parity plot between predicted and measured FR with deployment guard on. Three results structure the reading. First, the deployment guard is a precondition for neural surrogates. When evaluated raw, both networks fail dramatically (PGNN RMSE 122.3 min, with bias inverted to +46.1 min and the worst-case over-prediction rising to 325.0 min; control RMSE 254.7 min, bias +148.1 min). A yield strength of 591 MPa is a $\sim+10\sigma$ input in standardised space, where a neural network extrapolates as an unconstrained linear map. The guard — which clips inputs to the training range, and which tree ensembles apply implicitly through leaf saturation — restores the PGNN to RMSE 56.7 min and re-establishes its conservative bias (−36.9 min). The raw-versus-clipped contrast is itself a practical finding: a deployed neural fire-resistance surrogate must ship with an input-validity guard, and only guarded comparisons against tree ensembles are fair.

Second, engineering safety behaviour transfers to real beams. Under the guard, the PGNN over-predicts the measured fire resistance of 13 of the 50 real specimens, against 18 for the architecture-matched control and 23 for XGBoost. Likewise, its guarded worst-case unsafe error is 31.7 min, against the control's 210.2 min — a 6.6-fold improvement in the single most safety-critical statistic in this study, achieved on data drawn from a distribution the model never saw during training. The property engineered in simulation — a conservative bias, a suppressed unsafe-error rate, and a bounded worst case — is precisely the property that survives the transfer to real data.

Third, XGBoost's apparent robustness is an artefact, not genuine accuracy. Its aggregate RMSE on the real tests (50.2 min) is the lowest of all models, but in the severe stratum, containing 40 of the 50 beams, its R^2 is −0.01 — indistinguishable from simply predicting the training-mean fire re-

sistance for every beam. An internal signature of leaf saturation confirms this: XGBoost is more accurate under severe extrapolation (RMSE 42.5 min) than under moderate extrapolation (75.8 min, where all seven beams were under-predicted by a mean of 68.8 min). This pattern is impossible for a model that genuinely responds to its inputs, and natural for one that is collapsing toward a constant. A constant predictor would achieve a low RMSE on this set while conveying no information about any individual design — and this is the same model that violated monotonicity on 55% of in-distribution beams. Aggregate RMSE should not be read as validated accuracy. Similarly, in the moderate stratum, where inputs lie outside the envelope but close enough that a responsive model could succeed, the ranking inverts decisively. Only the physics-constrained variants achieve positive R^2 in any severity stratum (PGNN 0.38, physics-only 0.34, RMSE $\approx$ 40–41 min), against -0.39 for control and -1.26 for XGBoost. With $n = 7$, we present this as a graded trend rather than a definitive ranking, but its direction is exactly what the proposed mechanism implies: that physical structure remains valid where data does not reach. (Mild-stratum metrics, $n = 3$, with nearly equal measured values, are tabulated for completeness but are uninformative, since the R^2 denominator degenerates.)

One confound in the moderate stratum should be stated explicitly. Its seven beams have measured fire resistances spanning 77–240 min, with a mean of 182 min against a corpus mean of 102 min, so "moderate extrapolation" also means "unusually long-lasting beams" in this sample. This does not weaken the leaf-saturation reading; it is the mechanism behind it, since a model collapsing toward the training mean must under-predict a set whose true values sit far above that mean. It does, however, mean that the stratum is not a clean isolation of extrapolation severity alone, and the positive R^2 attained there by the physics-constrained variants should be read with that in mind.

3.9. Synthesis

Taken together, the five phases answer the three deficiencies posed in the problem statement. Against physical inconsistencies, the physics constraints alone cut the violation rate from 55–68% to 16.4% at zero measured accuracy cost; combined with the asymmetric objective they reach 6.6% in-distribution and 0.2% beyond it, at a cost of approximately 0.07 in R^2. Against uncontrolled extrapolation, the constrained model fails conservatively, with a bounded worst case and a roughly sevenfold tighter run-to-run dispersion in

out-of-distribution safety loss. The constrained models also retained predictive skill under moderate real-world extrapolation, uniquely among the models tested. Against risk-neutral training, the asymmetric objective converts an implicit design margin into an explicit, learned one, cutting unsafe predictions by nearly 40% in-distribution and by over 60% out-of-distribution, in exchange for a quantified and decomposable R^2 cost. The study's honest negative result is equally clear: no surrogate architecture evaluated here demonstrates predictive skill in the severe extrapolation regime, which constitutes 80% of the available real tests. This finding suggests that the primary limitation lies in the training dataset rather than the modelling framework. The material-property ranges explored in the simulation campaign are considerably narrower than those encountered in real construction practice. Expanding the dataset — in particular by increasing the variability in f_y, f_c, and insulation depth h_i — is essential for developing surrogate models that are not only robust under numerical stress testing but also capable of reliable experimental validation. The trade-off between conservatism and predictive accuracy across values of the asymmetry parameter, previously identified as an open question, is resolved in Section 3.6: the exchange rate is constant up to $\alpha = 0.8$ and halves beyond it, placing the adopted value at the knee of the frontier. Future research should prioritise broadening the material-property space and extending the multi-seed evaluation framework to split-level variance.

4. Conclusion

This study presents the design and evaluation of a physics-guided neural network (PGNN) for estimating the fire resistance of insulated FRP-strengthened RC beams. The framework integrates gradient-based monotonicity enforcement, bound constraints, collocation at the limit state beyond the training data envelope, and an asymmetric safety-oriented loss function. Drawing on a five-seed factorial ablation study benchmarked against ensemble and neural baselines, a model-agnostic assessment of physical consistency, and blind validation using 50 real furnace tests, the following conclusions emerge:

1. **Enforcing physics does not sacrifice accuracy.** Over five random seeds, the physics-constrained network performed on par with

its unconstrained counterpart of identical architecture in terms of in-distribution accuracy ($R^2 = 0.877 \pm 0.005$ versus 0.870 ± 0.014). Incorporating domain knowledge came at no measurable cost, and the expected trade-off between accuracy and physical consistency failed to appear.

2. **Unconstrained models are widely inconsistent with physics.** When subjected to finite-perturbation tests, the unconstrained MLP generated physically implausible outputs for 67.7% of the test beams, while XGBoost — the strongest baseline in terms of accuracy — did so for 55.3%. The PGNN reduced these violation rates to 6.6% in-distribution and just 0.2% out-of-distribution, conditions under which the unconstrained models still exceeded 40%.
3. **The asymmetric loss quantifies conservative behaviour.** Applying a 4:1 penalty for over-prediction relative to under-prediction builds in roughly a 12-minute safety margin, lowering the rate of unsafe predictions by 40% in-distribution (49% to 30%) and by over 60% out-of-distribution (40% to 16%). A six-level sweep of the asymmetry parameter establishes that this penalty ratio sits at the knee of the conservatism–accuracy frontier: up to $\alpha = 0.8$ each 0.01 of R^2 surrendered removes a steady 2.5 percentage points of unsafe prediction, and beyond it that return halves. The same sweep shows the benefit of the asymmetric objective to be entirely an extrapolation effect, since in-distribution safety loss is flat in α over this range.
4. **Constraints improve reliability across training runs.** The imposition of PGNN constraints cut the seed-to-seed dispersion of out-of-distribution safety loss by roughly a factor of seven (± 15.0 versus ± 106.8 min^2), with a smaller, complementary reduction in the dispersion of the worst-case unsafe error (± 15.6 versus ± 36.9 min) — which itself fell to 80.8 ± 15.6 minutes against 143.8 ± 36.9 minutes for the unconstrained control. This indicates that the constraints stabilise the behaviour of the entire model class rather than depending on a single favourable training instance — a characteristic of direct relevance for certification purposes.
5. **Safety-relevant behaviour, rather than raw accuracy, transfers from simulation to real-world data.** With an input-clipping safeguard applied at deployment — demonstrated here to be essential for any neural surrogate — the PGNN preserved its conservative tendency when evaluated against real furnace-test data, producing the

fewest unsafe predictions (13 of 50) and a guarded worst-case unsafe error 6.6 times lower than that of the unconstrained control (31.7 min versus 210.2 min). The seemingly strong aggregate error of XGBoost was found to stem from leaf saturation toward the training-set mean ($R^2 \approx 0$) rather than genuine predictive capability; in the only data stratum where predictive skill could be meaningfully assessed, only the physics-constrained variants attained positive R^2.

6. **The main constraint lies in the data, not the modelling approach.** None of the surrogate models, regardless of architecture, showed genuine predictive capability for the 80% of real tests that lay more than three training standard deviations outside the training envelope. This stems from the narrow material-property ranges used in the simulation campaign ($f_y = 414$–460 MPa; $f_c = 27$–40 MPa), which are considerably narrower than those found in laboratory and field practice. Broadening these ranges is a necessary precondition for developing experimentally validated surrogate models of any kind.

Future research should pursue three directions: regenerating the simulation dataset across realistic material-property ranges, identified here as the primary limitation; extending the asymmetry sweep of Section 3.6 into a full $\alpha\times$ physics grid, so that the two contributions to the seed-dispersion collapse can be separated at every asymmetry level rather than only at the two levels covered by the factorial ablation; and extending the replication methodology to include split-level variance analysis and larger ensembles of random seeds. The evaluation methods introduced in this work — including the Physics Violation Rate, directional safety metrics, seed-dispersion reporting, and severity-graded blind validation — have applicability to safety-critical surrogate modelling extending well beyond the present study.

References

[1] C.-T. Hsu, W. Punurai, H. Bian, Y. Jia, Flexural strengthening of reinforced concrete beams using carbon fibre reinforced polymer strips, Magazine of Concrete Research 55 (3) (2003) 279–288.

[2] G. Spadea, F. Bencardino, R. Swamy, Structural behavior of composite rc beams with externally bonded cfrp, Journal of Composites for Construction 2 (3) (1998) 132–137.

[3] J. Bonacci, M. Maalej, Behavioral trends of rc beams strengthened with externally bonded frp, Journal of Composites for Construction 5 (2) (2001) 102–113.

[4] L. A. Bisby, M. F. Green, V. K. Kodur, Response to fire of concrete structures that incorporate frp, Progress in structural engineering and materials 7 (3) (2005) 136–149.

[5] A. Ahmed, V. K. R. Kodur, The experimental behavior of frp-strengthened rc beams subjected to design fire exposure, Engineering Structures 33 (7) (2011) 2201–2211. doi:10.1016/j.engstruct.2011.03.010.

[6] P. P. Bhatt, V. K. R. Kodur, M. Z. Naser, Dataset on fire resistance analysis of frp-strengthened concrete beams, Data in Brief 52 (2024) 110031. doi:10.1016/j.dib.2024.110031.

[7] V. K. R. Kodur, A. Ahmed, Numerical model for tracing the response of frp-strengthened rc beams exposed to fire, Journal of Composites for Construction 14 (6) (2010) 730–742. doi:10.1061/(ASCE)CC.1943-5614.0000129.

[8] J.-G. Dai, W.-Y. Gao, J. G. Teng, Finite element modeling of insulated frp-strengthened rc beams exposed to fire, Journal of Composites for Construction 19 (2) (2015) 04014046. doi:10.1061/(ASCE)CC.1943-5614.0000509.

[9] Z. Ye, S.-C. Hsu, H.-H. Wei, Real-time prediction of structural fire responses: A finite element-based machine-learning approach, Automation in Construction 136 (2022) 104165. doi:10.1016/j.autcon.2022.104165.

[10] A. J. Keane, I. I. Voutchkov, Robust design optimization using surrogate models, Journal of Computational Design and Engineering 7 (1) (2020) 44–55. doi:10.1093/jcde/qwaa005.

[11] M. Z. Naser, V. K. R. Kodur, Explainable machine learning using real, synthetic and augmented fire tests to predict fire resistance and spalling of rc columns, Engineering Structures 253 (2022) 113824. doi:10.1016/j.engstruct.2021.113824.

[12] H. Kumarawadu, P. Weerasinghe, J. S. Perera, Evaluating the performance of ensemble machine learning algorithms over traditional machine

learning algorithms for predicting fire resistance in frp strengthened concrete beams, Electronic Journal of Structural Engineering 24 (3) (2024) 47–53.

[13] T. N.-T. Ho, T.-P. Nguyen, G. T. Truong, Concrete spalling identification and fire resistance prediction for fired rc columns using machine learning-based approaches: Concrete spalling identification and fire resistance prediction for fired rc columns, Fire Technology 60 (3) (2024) 1823–1866.

[14] P. Bhatt, N. Sharma, V. Kodur, M. Naser, Machine learning approach for predicting fire resistance of frp-strengthened concrete beams, Structural concrete 26 (4) (2025) 4143–4165.

[15] S. Wang, Y. Fu, S. Ban, Z. Duan, J. Su, Genetic evolutionary deep learning for fire resistance analysis in fibre-reinforced polymers strengthened reinforced concrete beams, Engineering Failure Analysis 169 (2025) 109149.

[16] S.-M. Kang, J.-K. Kim, Prediction of the moment capacity of frp-strengthened rc beams exposed to fire using anns, KSCE Journal of Civil Engineering 27 (8) (2023) 3471–3483.

[17] S.-Y. Zhang, S.-Z. Chen, X. Jiang, W.-S. Han, Data-driven prediction of frp strengthened reinforced concrete beam capacity based on interpretable ensemble learning algorithms, in: Structures, Vol. 43, Elsevier, 2022, pp. 860–877.

[18] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, V. Kumar, Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Transactions on Knowledge and Data Engineering 29 (10) (2017) 2318–2331. doi:10.1109/TKDE.2017.2720168.

[19] X. Jia, J. Willard, A. Karpatne, J. S. Read, J. A. Zwart, M. Steinbach, V. Kumar, Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles, ACM Transactions on Data Science 2 (3) (2021) 20. doi:10.1145/3447814.

[20] S.-M. Kang, C.-Y. Lee, J.-K. Kim, Ann based fire resistance prediction of frp-strengthened rc slabs with fireproof panel including air layer, Journal of Building Engineering 91 (2024) 109512.

[21] D. Runje, S. M. Shankaranarayana, Constrained monotonic neural networks, in: International Conference on Machine Learning, PMLR, 2023, pp. 29338–29353.

[22] K. Xu, M. Zhang, J. Li, S. S. Du, K.-i. Kawarabayashi, S. Jegelka, How neural networks extrapolate: From feedforward to graph neural networks, in: International Conference on Learning Representations, 2021. doi:10.48550/arXiv.2009.11848.

[23] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[24] H. Zhang, D. Nettleton, Z. Zhu, Regression-enhanced random forests, arXiv preprint arXiv:1904.10416 (2019).

[25] A. J. Patton, A. Timmermann, Properties of optimal forecasts under asymmetric loss and nonlinearity, Journal of Econometrics 140 (2) (2007) 884–918.

[26] M. Raissi, P. Perdikaris, G. 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 (2019) 686–707.

[27] H. Hu, L. Qi, X. Chao, Physics-informed neural networks (pinn) for computational solid mechanics: Numerical frameworks and applications, Thin-Walled Structures 205 (2024) 112495.

[28] R. Yarmohammadian, F. Put, R. Van Coile, Physics-informed surrogate modelling in fire safety engineering: a systematic review, Applied Sciences 15 (15) (2025) 8740.

[29] J. Monteiro, M. O. Ahmed, H. Hajimirsadeghi, G. Mori, Monotonicity regularization: Improved penalties and novel applications to disentangled representation learning and robust classification, in: Uncertainty in Artificial Intelligence, PMLR, 2022, pp. 1381–1391.

[30] M. Gupta, A. Cotter, J. Pfeifer, K. Voevodski, K. Canini, A. Mangylov, W. Moczydlowski, A. Van Esbroeck, Monotonic calibrated interpolated look-up tables, Journal of Machine Learning Research 17 (109) (2016) 1–47.

[31] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[32] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).

[33] P. Ramachandran, B. Zoph, Q. V. Le, Searching for activation functions, arXiv preprint arXiv:1710.05941 (2017).

[34] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.

[35] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[36] N. Poudel, S. Kandel, Source code for a physics-guided neural network (pgnn) for predicting the fire resistance of insulated frp-strengthened rc beams (Aug. 2026). doi:10.5281/zenodo.21982861.
URL `https://doi.org/10.5281/zenodo.21982861`

[37] T. Hastie, R. Tibshirani, J. Friedman, et al., The elements of statistical learning (2009).