UNBIASED PHYSICS LOSSES UNDER MODEL UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-informed machine learning usually penalizes violations of a chosen governing equation. However, when that equation is misspecified, minimizing the nominal residual can instead increase the error in the predicted solution. To make physics-informed machine learning effective under such misspecification, we propose compatibility training, which uses the smallest residual over a plausible family of equations. At each training step, compatibility training selects the equation in this family that best fits the current prediction and penalizes only deviations from that equation. Our theory establishes a mismatch-dependent bias floor for nominal training and characterizes the physics constraints that remain after the uncertain parameters are adjusted. We derive an exact local curvature identity that shows when the retained physics constraints and observations together identify the true solution. This curvature identity connects solution recovery to the structure and extent of the uncertainty family. Across multiple synthetic PINN benchmarks, PDEBench trajectory and operator tasks, and Darcy operator-learning experiments, compatibility training improves prediction under model mismatch, including in settings with sparse or noisy observations and field-valued uncertainty. In field-valued Darcy flow, compatibility training also reduces the true residual by more than an order of magnitude.

1 INTRODUCTION

Physics-informed machine learning trains predictors not only to fit observations but also to satisfy governing equations. Physics-informed neural networks (PINNs) and physics-informed neural operators implement this principle by adding a residual penalty to the task loss (Raissi et al., 2019; Li et al., 2024),

$$\mathcal{L}(\theta) \coloneqq \mathcal{L}_{\mathrm{data}}(f_\theta) + \lambda \|R_{\hat{\mu}}(f_\theta)\|^2.$$

When $R_{\hat{\mu}}$ encodes the true governing equation, minimizing this residual encourages predictions to obey the correct physics.

In scientific workflows, however, the equation encoded in the loss can be wrong: A coefficient in the equation may be measured inaccurately or reused in a new physical setting where the correct coefficient value is different. In particular, coefficients that summarize unresolved or fine-scale physics can require different values at different modeling scales. In Darcy flow, permeability is a coefficient in the governing equation; replacing a spatially varying or sample-specific field with one constant makes the loss enforce an equation that omits variations affecting the true solution. Such mismatch is a standard concern in computer-model calibration, closure modeling, and imperfect-model data assimilation (Kennedy & O'Hagan, 2001; Duraisamy et al., 2019; Tr'emolet, 2006). Under such mismatch, minimizing the nominal residual can push the predictor away from the true solution: The training residual may decrease even as prediction error relative to the true solution grows. Because the nominal residual is both an optimized loss term and a routinely inspected training diagnostic (Wang et al., 2021), a training run that is failing at prediction can give a misleading appearance of success.

Our approach represents uncertainty about the governing equation through a family of physically plausible alternatives. Measurements and physical knowledge constrain the allowed coefficients and spatial corrections, and each allowed choice defines one equation in the family. This setting occurs

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.

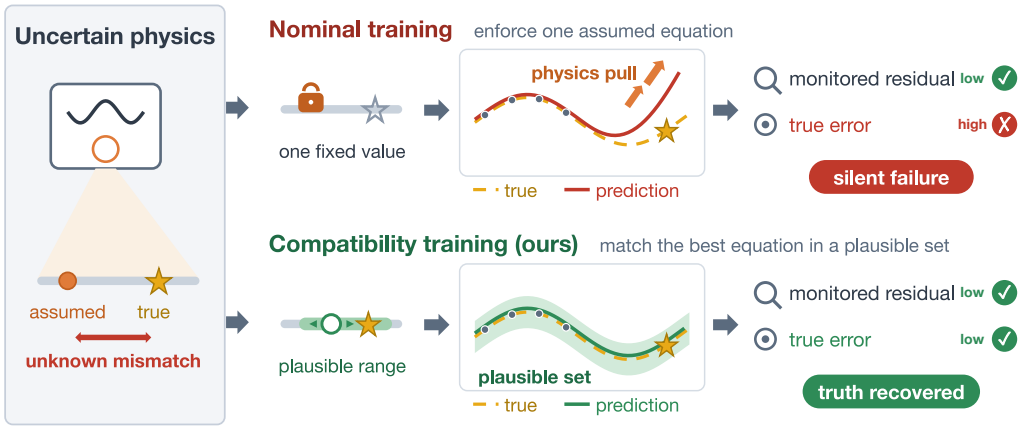


Figure 1: Conceptual overview of physics-informed machine learning under model uncertainty. The standard practice of nominal training penalizes the residual of one assumed governing equation. When that equation is wrong, training updates driven by the nominal residual loss can lower the monitored residual while moving the prediction away from the true solution. Our proposed compatibility training selects the equation in a plausible family that best fits the current prediction and penalizes the residual of that equation. When the family contains the true equation, a prediction equal to the true solution receives no physics penalty; predictions that violate every equation in the family still receive a penalty.

in turbulence modeling, where physically constrained Reynolds-stress corrections are calibrated on observed flows and sampled to predict related flows (Wu et al., 2016). Fig. 1 summarizes the failure of nominal training and our proposed compatibility training.

This paper therefore asks which physics loss a scientific machine learning (SciML) model should use when the governing equation is uncertain, and how to optimize that loss. Our setup takes this equation family as an input and collects the allowed parameter choices in $\mathcal{U}$. Measurement tolerances provide bounds on uncertain coefficients (Milanese & Vicino, 1991). Physical constraints or separate calibration data guide the choice of spatial correction patterns (Wu et al., 2016). Our compatibility objective asks whether a prediction is consistent with at least one equation in that family:

$$\mathcal{J}_{\mathrm{comp}}(u) := \mathcal{L}_{\mathrm{data}}(u) + \lambda \mathcal{C}_{\mathcal{U}}(u), \qquad \mathcal{C}_{\mathcal{U}}(u) := \min_{\mu \in \mathcal{U}} \|R_\mu(u)\|^2.$$

The residual loss supplies the physics supervision used to train the predictor. This training signal should penalize violations that the allowed changes to the equation cannot explain. At each training step, compatibility training profiles the uncertain physics parameters by minimizing the current prediction's residual over the allowed parameter choices. Compatibility training then penalizes only the remaining unexplained violations when updating the predictor. When the family contains the true equation, a prediction equal to the true solution receives no physics penalty. Restricting the family to the single nominal equation recovers standard nominal training. The selected physics parameter is a nuisance variable (Murphy & Van der Vaart, 2000): The nuisance parameter is optimized to evaluate the compatibility loss, while the prediction is the quantity we aim to recover.

The allowed parameter choices generate an uncertainty family of equations. This uncertainty family determines which model discrepancies can be absorbed and which physics constraints remain useful for prediction. Profiling adjusts the uncertain physics parameters to explain the prediction's residual and penalizes what remains unexplained (Theorem 4). These remaining constraints can distinguish predictions that the observations alone leave unresolved (Corollary 5). In compatibility training, the distinction between absorbed mismatch and remaining physics constraints guides the choice of uncertain coefficients, spatial correction modes, and parameter ranges: Cover plausible model errors while retaining constraints that distinguish predictions fitting the same data (Sec. 4).

With the same parameter family and loss, compatibility training is joint inverse training after the physical parameters have been optimized out. Optimizing out the physical parameters is known as variable projection. Nuisance elimination is classical in separable least squares (Golub & Pereyra,

1973; 2003) and neural-network training (Newman et al., 2021); physical-parameter elimination has also been used for inverse problems for partial differential equations (PDEs) (Dong & Wang, 2023). We use this connection to explain how a wrong equation biases prediction and how allowing additional correction directions changes the physics information available to the predictor (Theorem 1 and Corollary 7).

Model-discrepancy learning is a closely related approach that adds a learned correction term to the governing equation to account for missing or inaccurate physics. Zou et al. (2024) augment PINNs with a correction network to improve solutions and models under sparse or noisy observations. This connection makes the choice of correction family central: The same flexibility that absorbs model error can also remove physics constraints needed to distinguish predictions. Our analysis quantifies this tradeoff and shows when the remaining constraints complement the observations. App. B details the connections to inverse PINNs, variable projection, and learned model-discrepancy corrections.

Contributions. First, we identify why standard nominal training can fail when the governing equation is wrong: Reducing the nominal residual loss can increase prediction error while giving the appearance of successful training (Sec. 2). Second, we propose a compatibility training framework that selects the best-fitting equation in a specified family at each training step and updates the predictor using the remaining residual (Sec. 3.1). Third, our theory proves a mismatch-dependent bias floor for nominal training and establishes when compatibility training recovers the true solution locally. An exact curvature identity shows how the retained physics constraints and observations jointly identify the solution and quantifies the information lost when new uncertainty directions are added (Secs. 3.2 and 4). We also analyze local tangent approximations and finite collocation. Fourth, experiments on synthetic PINNs, PDEBench trajectories and operators (Takamoto et al., 2022), and Darcy Fourier neural operators (FNOs) with scalar and field-valued uncertainty (Li et al., 2021) show how compatibility training improves prediction under mismatch, how the physics constraints that remain after profiling help prediction with sparse or noisy observations, and why matching the uncertainty family matters (Sec. 5).

2 Misspecified Physics Bias

Let $R_\mu(u)$ denote the physics residual. The true solution $u^\star$ satisfies the governing equation $R_{\mu^\star}(u^\star) = 0$, where $\mu^\star$ is the true physics parameter. We study the setting in which training uses a mismatched nominal parameter $\hat{\mu} = \mu^\star + \Delta_\mu$. Nominal training minimizes

$$\mathcal{J}_{\mathrm{nom}}(u) \coloneqq \mathcal{L}_{\mathrm{data}}(u) + \lambda P_{\hat{\mu}}(u), \qquad P_{\hat{\mu}}(u) \coloneqq \|R_{\hat{\mu}}(u)\|^2.$$

The problem is that parameter mismatch can give the nominal physics loss a nonzero gradient even when the prediction already matches the true solution. Specifically, a small mismatch in the nominal physics parameter changes the residual to first order through the parameter derivative, or nuisance Jacobian. Lemma 1 in App. C gives

$$\nabla_u P_{\hat{\mu}}(u^\star) = 2 D_u R_{\mu^\star}(u^\star)^* \partial_\mu R_{\mu^\star}(u^\star) \Delta_\mu + O(\|\Delta_\mu\|^2).$$

This nominal physics gradient is generically nonzero. When the nominal physics gradient is nonzero, a gradient step on the nominal physics loss moves an already correct prediction away from the true solution $u^\star$.

Theorem 1 (Stationary-point bias floor for nominal training)**.** *Let $\mathcal{N}$ be an open convex neighborhood of $u^\star$ and let $\lambda > 0$. Suppose $\mathcal{L}_{\mathrm{data}}$ and $P_{\hat{\mu}}$ are Fréchet differentiable on $\mathcal{N}$, with $\nabla \mathcal{L}_{\mathrm{data}}(u^\star) = 0$, and the gradients of $\mathcal{L}_{\mathrm{data}}$ and $P_{\hat{\mu}}$ are respectively M- and K-Lipschitz, where $M > 0$ and $K \geq 0$. If $g \coloneqq \|\nabla_u P_{\hat{\mu}}(u^\star)\| > 0$, every stationary point $u_{\lambda,\hat{\mu}} \in \mathcal{N}$ satisfies*

$$\|u_{\lambda,\hat{\mu}} - u^\star\| \geq \frac{\lambda g}{M + \lambda K}. \tag{1}$$

The right-hand side is strictly increasing in λ.

Here g is the magnitude of the nominal physics-loss gradient at the true solution. This magnitude scales linearly with the parameter mismatch, $g = \Theta(\|\hat{\mu} - \mu^\star\|)$, when the bounds in Lemma 1 and Corollary 1 hold uniformly near $\mu^\star$. Thus nominal training has a first-order mismatch bias that tuning a positive physics weight cannot remove. App. C gives the matching local upper bound, complete corollaries, and the bias-sandwich interpretation.

3 Compatibility Residuals and Retained Physics Constraints

This section explains when compatibility training can remove the bias caused by a wrong equation while keeping physics constraints useful for predicting the solution. We define our proposed training loss, identify the constraints that remain after adjusting uncertain parameters, and show how these constraints work with observations to recover the solution.

3.1 Global Profile

For a compact set $\mathcal{U}$ of allowed physics parameters, define the compatibility loss $\mathcal{C}_{\mathcal{U}}$ as the smallest squared residual over the corresponding equations, and the full training objective $\mathcal{J}_{\mathrm{comp}}$ by adding the weighted compatibility loss to the data term:

$$\mathcal{C}_{\mathcal{U}}(u) \coloneqq \min_{\mu \in \mathcal{U}} \|R_\mu(u)\|^2, \qquad \mathcal{J}_{\mathrm{comp}}(u) \coloneqq \mathcal{L}_{\mathrm{data}}(u) + \lambda \mathcal{C}_{\mathcal{U}}(u).$$

For a fixed prediction u, we first choose the equation in the family that minimizes the residual for u. We then use the remaining residual to update the predictor, together with the data loss. If the family contains the true equation ($\mu^\star \in \mathcal{U}$), the true solution satisfies an allowed equation and therefore receives no physics penalty: $\mathcal{C}_{\mathcal{U}}(u^\star) = 0$.

Restricting the uncertainty set to $\mathcal{U} = \{\hat{\mu}\}$ gives $\mathcal{C}_{\mathcal{U}}(u) = \|R_{\hat{\mu}}(u)\|^2$ and recovers nominal training.

Compatibility training and joint inverse training seek the same globally optimal predictors under the matched setting of Proposition 2 in App. D.2. Profiling solves for the best physics parameter at each current prediction, while joint gradient training updates the predictor and a persistent physics parameter together. Optimizing out that parameter changes the optimization geometry seen by the predictor, so the training paths can differ even though both objectives have the same set of globally optimal predictors. App. D.2 states the exact relation between the minimizers and extends the minimizer relation to instance-specific nuisance families. We next identify which physics constraints still distinguish predictions after the uncertain physics parameters have been optimized out.

3.2 Retained Curvature and Identifiability

The remaining physics constraints are useful if the constraints help distinguish predictions that fit the same observations. To make the role of these constraints precise, consider a compatible prediction–parameter pair $(u^\star, \mu^\star)$, which satisfies $R_{\mu^\star}(u^\star) = 0$. We define the prediction derivative A, the physics-parameter derivative J, and the projector Q that removes residual changes caused by the uncertain parameters:

$$A \coloneqq D_u R_{\mu^\star}(u^\star), \qquad J \coloneqq \partial_\mu R_{\mu^\star}(u^\star), \qquad Q \coloneqq I - \Pi_{\mathrm{Range}(J)}.$$

Here Ah and Jv are the first-order residual changes caused by a prediction perturbation h and a parameter perturbation v, respectively. The projector Q keeps the component orthogonal to $\mathrm{Range}(J)$, which contains the residual directions that parameter changes can explain. In the local linearized model, allowing the physics parameter to adjust gives

$$\min_{v \in \mathbb{R}^p} \|Ah + Jv\|^2 = \|QAh\|^2. \tag{2}$$

In this local model, $QJ = 0$ means that Q removes every first-order residual change caused by adjusting the physics parameters. Thus Q removes the first-order nominal residual mismatch $J\Delta_\mu$ from Sec. 2, while QAh retains the residual response that uncertain physics cannot explain. Our central geometric identity gives the exact profiled Hessian under the regular interior profile hypotheses of Theorem 4:

$$\nabla \mathcal{C}_{\mathcal{U}}(u^\star) = 0, \qquad D^2 \mathcal{C}_{\mathcal{U}}(u^\star)[h,k] = 2\langle QAh, QAk \rangle. \tag{3}$$

In a prediction direction h, the retained physics curvature $2\|QAh\|^2$ is the second derivative of the profiled physics loss at $u^\star$ along h. This curvature measures how strongly the loss penalizes a prediction change along h after the physics parameters are optimized again to minimize the residual for the changed prediction. If the observations cannot detect a prediction change in direction h but $QAh \neq 0$, the physics loss still penalizes that change. These remaining physics constraints can complement incomplete observations. The inequality below requires the combined data–physics

loss to penalize every prediction direction, with a uniform positive lower bound on curvature. If, for some $\gamma > 0$,

$$D^2\mathcal{L}_{\mathrm{data}}(u^\star)[h,h] + 2\lambda\|QAh\|^2 \geq \gamma\|h\|^2 \qquad (h \in \mathcal{H}),$$

and the data gradient vanishes at $u^\star$, then $u^\star$ uniquely minimizes $\mathcal{J}_{\mathrm{comp}}$ among all predictions u in a sufficiently small ball centered at $u^\star$ (Corollary 5). Local recovery of the true solution $u^\star$ depends on the combined information from data and physics, so the observations alone need not distinguish the true solution from every nearby alternative. In finite dimensions with data Hessian O^*O, the condition is equivalent to $\ker O \cap \ker(QA) = \{0\}$: Every nonzero prediction change invisible to the data must be detected by the constraints that remain after profiling.

4 Uncertainty-Family Design

An uncertainty family is a set of plausible governing equations obtained by varying uncertain coefficients or spatial corrections. The family specifies how compatibility training is allowed to adjust the governing equation to explain a prediction's residual. The goal is to cover plausible equation errors while keeping physics constraints that distinguish predictions fitting the same observations. We now explain how to choose the allowed equation changes and parameter bounds. Computing and approximating the resulting loss are discussed in App. D.8.

4.1 Choosing Uncertainty Directions and Bounds

We construct an uncertainty family for compatibility training by specifying which coefficients in $R_\mu(u)$, or which spatial correction modes, may vary. The family also specifies which observations share the same physics parameter: For example, one coefficient may be shared across an entire trajectory, or each input function may have a separate field correction. Measurement tolerances, physical bounds, or separate discrepancy samples guide the parameter ranges. For a positive scalar coefficient in the governing equation, such as viscosity or a reaction rate, a multiplicative uncertainty interval takes the form $\mathcal{U} = [\hat{\mu}/\rho, \rho\hat{\mu}]$ with $\hat{\mu} > 0$ and $\rho \geq 1$. This interval allows the coefficient to decrease or increase by up to a factor of ρ from the coefficient's nominal value $\hat{\mu}$. For a coefficient field such as Darcy permeability, prescribed or independently learned spatial basis functions define the allowed correction patterns. Bounds on the basis coefficients limit the size of these corrections. The basis, normalization, parameter bounds, and sharing rule are fixed before the trained predictor is evaluated on test data; the profiled coefficients still change with the predictor during training.

Here nuisance directions are the first-order residual changes that adjusting these uncertain physics parameters can produce. Adding such directions lets profiling absorb more equation error, but the added nuisance directions can also remove penalties on incorrect predictions. For nested nuisance tangent spaces $\mathcal{T}_1 \subseteq \mathcal{T}_2$ at the same compatible solution with the same prediction derivative A, the reduction in physics curvature along a prediction perturbation h is $2\|\Pi_{\mathcal{E}}Ah\|^2$, where $\mathcal{E} = \mathcal{T}_2 \cap \mathcal{T}_1^\perp$ contains the newly added residual directions. The extra adjustments can now explain away this component of the residual change Ah, weakening the penalty on the prediction change h. Corollaries 7 and 8 state the assumptions for this identity and when the remaining physics constraints and observations still identify the solution.

Enlarging the uncertainty radius changes how far the existing corrections may move. Increasing ρ widens the multiplicative uncertainty interval $[\hat{\mu}/\rho, \rho\hat{\mu}]$ for the scalar coefficient; increasing the bound on the field correction coefficients allows larger corrections in the same spatial basis. This increase in the uncertainty radius can bring a previously excluded true equation into the family. At a fixed interior point of the same parameterization, changing the radius does not add residual tangent directions. Thus, even a larger radius cannot supply a spatial mode missing from the fixed basis.

Prediction error on held-out observations guides the choice among candidate families. If $\mathcal{U}_1 \subseteq \mathcal{U}_2$, the second family allows every parameter choice in the first family and possibly more. Consequently, $\mathcal{C}_{\mathcal{U}_2}(u) \leq \mathcal{C}_{\mathcal{U}_1}(u)$: For the same prediction u, more parameter choices can only lower or leave unchanged the minimum residual. A larger family can therefore produce a smaller residual without improving the prediction at all, as the comparison with fixed predictions in Table 30 demonstrates.

5 EXPERIMENTS

The experiments below test whether the proposed compatibility training improves prediction when the uncertainty family covers the equation error, whether the remaining physics constraints help with sparse or noisy observations, and how the allowed correction patterns and parameter ranges affect prediction.

5.1 EXPERIMENTAL SETUP

Metrics and mismatch design. Data are generated by a true system with parameter $\mu^\star$, while the physics loss is evaluated using a corrupted nominal parameter $\hat{\mu}$. We measure prediction with relative L^2 error (RelL2). To summarize performance across each mismatch grid, we define the graceful degradation score (GDS) as the mean prediction-error ratio to data-only:

$$\mathrm{RelL2} \coloneqq \frac{\|u_\theta - u^\star\|_2}{\|u^\star\|_2}, \qquad \mathrm{GDS} \coloneqq \frac{1}{N_\delta} \sum_{k=1}^{N_\delta} \frac{\mathrm{Err}_{\mathrm{meth}}(\delta_k)}{\mathrm{Err}_{\mathrm{data}}(\delta_k) + 10^{-12}},$$

where N_δ is the number of tested mismatch levels, and $\mathrm{Err}_{\mathrm{meth}}(\delta_k)$ and $\mathrm{Err}_{\mathrm{data}}(\delta_k)$ are the seed-averaged RelL2 values of the method and data-only baseline at mismatch δ_k, respectively. We also report empirical residual mean squared errors (MSEs), where a lower MSE means closer agreement with the equation used to evaluate that residual. The harm threshold is the smallest tested mismatch ratio δ for which a method's seed-averaged RelL2 exceeds the data-only error by more than 2%. Higher values indicate a later crossing as δ increases; ∞ means that the method stays within the error threshold at all tested ratios.

Benchmarks and baselines. The synthetic PINN benchmarks use periodic Burgers, Fisher–Kolmogorov–Petrovsky–Piskunov (Fisher–KPP) reaction–diffusion, and a non-affine reaction-law variant, with mismatch levels $\delta \in \{0.25, 1, 4, 16\}$. For Burgers, $\hat{\nu} = \delta\nu^\star$ changes the viscosity in the diffusion term νu_{xx}; for Fisher–KPP, $\hat{\alpha} = \delta\alpha^\star$ changes the reaction rate in $\alpha u(1-u)$. For the non-affine variant, $\hat{\beta} = \log\delta$ changes the exponential factor in the reaction term $\alpha e^{\beta u} u(1-u)$, with true $\beta^\star = 0$. Standardized-data checks use PDEBench Burgers trajectories and a three-channel compressible Navier–Stokes FNO. Darcy operator-learning experiments use FNOs with scalar coefficient corruption and sample-specific field corruption with prescribed or independently learned eight-mode bases.

We compare data-only, nominal and inverse physics, worst-case and uncertainty-average residuals, global compatibility training, and local tangent projection. These baselines compare fixed and jointly updated physics, worst-case and average uncertainty aggregation, and local approximation. For weight-tuned comparisons, every physics method in the comparison is trained at the same four candidate weights. We choose each method's weight by the lowest MSE on shared held-out validation observations, separately for each mismatch–seed setting and each trajectory when applicable. Fixed-weight comparisons use a common λ within each block. Table 6 summarizes the selection budgets; test error is never used for selection.

Operator loss and reporting. The field-valued Darcy loss uses an approximate profile and holds the computed coefficients fixed when differentiating the loss to update the predictor; positivity clipping in the generator can create coefficient changes that the chosen linear family cannot represent (App. F.5).

Within each comparison, compatibility training and all baselines share observation draws or data splits and predictor initialization seeds. Each configuration uses four seeds, except the public compressible-flow check, which uses two data-split seeds. For aggregate entries, we first average the listed mismatch levels and trajectories within each seed, then average across seeds. The $\pm$ term is one standard error of this final mean: the standard deviation of the seed averages divided by the square root of the number of seeds.

Computational cost. Recorded run time for compatibility training is $0.98\times$ to $1.24\times$ nominal PINN time in matched fixed-weight runs and $1.68\times$ nominal run time for eight-mode field Darcy. Predictions from the trained network do not require optimizing the physics parameters.

Table 1: Nominal PINN weight sweep for reaction–diffusion at mismatch 16. Lower RelL2 indicates better prediction. Boldface marks each column's minimum, including the nominal-residual minimum that disagrees with the prediction and true-residual minima.

λ	RelL2	True MSE	Nominal MSE
0.02	**0.5100**	**1.3223**	0.8434
0.05	0.5848	1.7396	0.1577
0.20	0.6239	2.0041	0.0105
1.00	0.6348	2.0588	$\mathbf{7 \times 10^{-4}}$

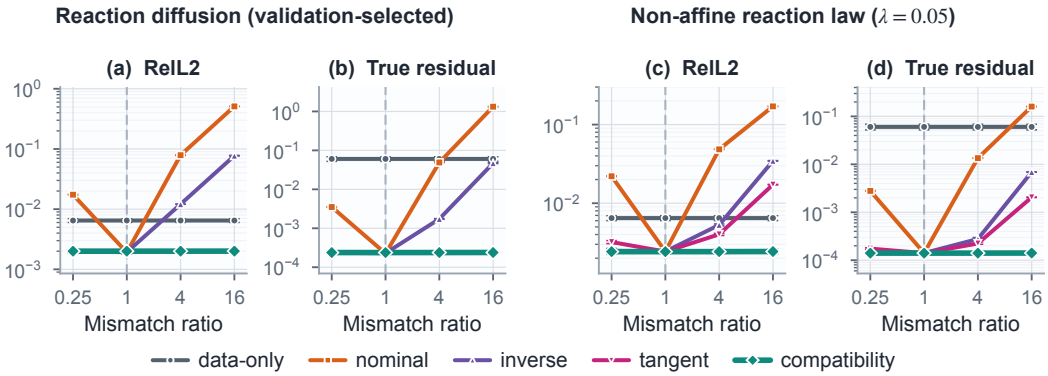


Figure 2: Prediction error and true residuals across mismatch levels, with matched weight-selection rules. The reaction–diffusion sweep uses validation-selected runs from Table 2, and the non-affine sweep uses $\lambda = 0.05$ for every physics method (Table 16). Bands show one standard error over four seeds, and the dashed guide marks the correctly specified point $\delta = 1$. Global profiling remains accurate as the local approximation degrades.

5.2 Nominal Residual Minimization Can Be Anti-Informative

The nominal weight sweep shows that a smaller training residual can accompany a larger prediction error. In Table 1, increasing λ reduces the nominal residual while increasing prediction error and the true residual. Selecting the smallest nominal residual therefore chooses the least accurate predictor in this sweep: A stronger penalty makes the prediction obey the wrong equation more closely.

This behavior is consistent with Theorem 1: Under the theorem's assumptions, the lower bound on prediction error at nominal stationary points grows with λ. Fig. 3 shows the predictions moving toward the wrong solution branch and evaluates the predictions under both the nominal and true equations. The non-affine benchmark in Table 18 exhibits the same behavior.

5.3 Removing Represented Mismatch

Main affine and non-affine mismatch sweeps. Fig. 2 shows how prediction error and true residuals vary with errors in the nominal physics parameters. At large mismatch, nominal and jointly trained inverse PINNs lose prediction accuracy, while compatibility training keeps prediction error and the true residual low. In the affine and non-affine PINN sweeps reported in Table 2 and Appendix Table 14, validation-tuned compatibility training lowers the reaction–diffusion GDS from 23.47 for tuned nominal physics and 3.60 for tuned inverse PINN to 0.31. In the non-affine benchmark, compatibility training reaches RelL2 0.00191 ± 0.00013, while data-only and tuned inverse PINN remain above 0.0064. With both global and tangent profiles fixed at $\lambda = 0.05$, the errors of global and tangent profiles are 0.00243 and 0.00673, respectively (Table 16). This gap shows the benefit of evaluating the nonlinear family directly when the family's local linearization becomes inaccurate.

Table 2: Main affine PINN results over four mismatch levels. Every physics row selects from the same four weights by held-out validation MSE. Table 15 includes worst-case training at a common weight. Lower RelL2 and GDS indicate better prediction. A higher Harm value indicates a later threshold crossing as δ increases; ∞ means no harm at any tested ratio. Boldface marks the lowest observed RelL2, true MSE, and GDS means within each benchmark. Values that round to the same minimum are all bolded.

Method	RelL2	True MSE	GDS	Harm
Burgers				
data-only	0.0114 ± 0.0015	0.4026 ± 0.0675	1.00	∞
nominal PINN, tuned	0.0144 ± 0.0003	0.0088 ± 0.0012	1.27	16
inverse PINN, tuned	0.0078 ± 0.0003	0.0040 ± 0.0004	0.69	16
uncertainty-average, tuned	0.0958 ± 0.0004	0.0942 ± 0.0030	8.42	1
compatibility training, tuned	$\mathbf{0.00254 \pm 0.00039}$	$\mathbf{0.00160 \pm 0.00018}$	**0.22**	∞
Reaction–diffusion				
data-only	0.00648 ± 0.00045	0.0604 ± 0.0080	1.00	∞
nominal PINN, tuned	0.1520 ± 0.0008	0.3437 ± 0.0091	23.47	0.25
inverse PINN, tuned	0.0233 ± 0.0002	0.0123 ± 0.0003	3.60	4
uncertainty-average, tuned	0.5781 ± 0.0007	0.1329 ± 0.0045	89.25	0.25
compatibility training, tuned	$\mathbf{0.00200 \pm 0.00012}$	$(\mathbf{2.38 \pm 0.13}) \times \mathbf{10^{-4}}$	**0.31**	∞

Table 3: PDEBench Burgers multi-trajectory check with a common fixed $\lambda = 0.2$ for every physics method. Separate neural fields use public trajectories $\{0, 1, 3\}$ and mismatch ratios $\{1, 16\}$; each row aggregates the same 24 trajectory–mismatch–seed cells. Observations and collocation use $t \leq 1$, with evaluation on $t \in [0, 2]$. Lower RelL2 and GDS indicate better prediction. Boldface marks each metric's lowest observed mean. Values that round to the same minimum are all bolded.

Method	RelL2	True MSE	GDS
data-only	0.2002 ± 0.0109	19.1 ± 1.6	1.000
nominal PINN	0.1772 ± 0.0033	1.87 ± 0.14	0.737
inverse PINN	0.1682 ± 0.0045	$\mathbf{1.56 \pm 0.14}$	0.703
compatibility training	$\mathbf{0.1511 \pm 0.0081}$	3.74 ± 0.27	**0.673**

5.4 Retained Physics Constraints under Weak Supervision

Sparse and noisy observations. Here we test whether the physics constraints retained by compatibility training improve prediction as observations become fewer or noisier. Under reaction–diffusion mismatch 16, compatibility training incurs no harm under the 2% criterion across 20 observation–noise settings, each evaluated with four matched seeds, and maintains seed-averaged RelL2 below 0.016 at 10% noise. Table 20 reports the 10% noise results. These results show that allowing coefficient corrections preserves useful physics supervision even when observations are few or noisy.

PDEBench trajectories. Across three public trajectories, compatibility training achieves the lowest average RelL2 and GDS at a common $\lambda = 0.2$ (Table 3), while nominal and inverse PINNs obtain lower true residuals. Prediction error and true-residual error therefore rank these predictors differently: Selecting the lowest true residual would miss the most accurate predictor in this comparison.

5.5 Matching the Uncertainty Family

Field-valued coefficient uncertainty. This comparison tests whether the allowed coefficient corrections match the spatial errors in the nominal inputs. A scalar profile cannot represent the sample-specific eight-mode coefficient error. At the common fixed weight in Table 4, compatibility training with the matching field profile reduces mean RelL2 by 11.5% relative to data-only. The paired mean difference is -0.00114, with pointwise 95% confidence interval (CI) $[-0.00151, -0.00078]$, and prediction improves in all eight matched mismatch–seed cells. Field compatibility training also reduces mean true-residual MSE by a factor of 13.8.

Table 4: Stronger field-valued Darcy FNO check with 1024 training functions, aggregated over the two harder mismatch amplitudes $\{0.5, 1.0\}$. All physics rows use the common fixed $\lambda = 10^{-5}$. Lower RelL2 and GDS indicate better prediction. Boldface marks each metric's lowest observed mean. Values that round to the same minimum are all bolded.

Method	RelL2	True MSE	Compat. MSE	GDS
data-only	0.00990 ± 0.00048	0.1355 ± 0.0062	0.1087	1.00
nominal physics	0.01373 ± 0.00029	0.0387 ± 0.0010	0.0132	1.37
scalar compatibility training	0.01292 ± 0.00028	0.0380 ± 0.0011	0.0148	1.29
field compatibility training	$\mathbf{0.00876 \pm 0.00040}$	$\mathbf{0.00983 \pm 0.00066}$	**0.0090**	**0.89**

Nominal training and scalar compatibility training reduce the true-residual diagnostic while worsening prediction. Allowing sample-specific spatial corrections turns an otherwise harmful physics term into a useful training signal in this setting. The approximate profiles also track the dominant coefficient mismatch more accurately than the same profiles applied post hoc to data-only predictions. Fig. 5 shows the spatial corrections and pressure errors for a test function selected by data-only error rather than the gain from compatibility training.

Constructing a basis without test corrections. In Darcy, eight modes are learned by principal component analysis (PCA) from 256 independent draws of a specified discrepancy prior and frozen before predictor training. At fixed radius $\rho = 1$, compatibility training reaches mean RelL2 0.06825, compared with 0.07121 for data-only and 0.06789 for nominal physics (Table 27). Thus a basis constructed independently of the test corrections can help prediction, although nominal physics gives a similar mean error in this setting. App. F.5 details the construction.

Uncertainty-set sensitivity. The radius sweep varies how much the coefficients in the governing equation may differ from the nominal coefficient values, while keeping the allowed correction patterns fixed. For reaction–diffusion, increasing ρ widens the allowed reaction-rate interval $[\hat{\alpha}/\rho, \rho\hat{\alpha}]$. At $\hat{\alpha} = 16$ and $\alpha^\star = 1$, radii 2 and 4 exclude the truth and compatibility training remains biased; radii 16 and 64 include the true reaction rate and both reach RelL2 0.0043 with true MSE 1.0×10^{-4} (Table 28). In this experiment, the coefficient bounds must therefore let the reaction rate decrease from the nominal value 16 to the true value 1. The fixed-basis Darcy sweep likewise benefits from greater coverage, without collapse at larger tested radii (Table 29). On fixed data-only Darcy predictions, increasing the post-hoc profile dimension from 4 to 16 lowers the profiled MSE from 0.6558 to 0.4983 with unchanged RelL2 (Table 30). Thus, the profile score depends on both prediction quality and the allowed correction family: Adding correction modes can lower the score without improving the prediction.

Further experimental details. App. F documents the full residual definitions, numerical approximations, training protocols, and consistency checks, together with the timing scope, baseline costs, and computational scaling. App. G reports paired differences and individual seed results, along with the remaining observation–noise results, additional benchmark comparisons, and coefficient-profile diagnostics.

6 Conclusion

We propose compatibility training for physics-informed machine learning when the governing equation used in training is wrong. The method adjusts uncertain physics parameters within a family of plausible equations to absorb model errors that the family can represent. Our proposed physics loss penalizes predictions that violate every equation in the family. Our analysis explains why enforcing a wrong equation biases prediction and how the constraints that remain after parameter adjustment help distinguish predictions that fit the same observations. Experiments show that compatibility training improves prediction with sparse or noisy observations. Prediction accuracy also depends on the allowed correction patterns and parameter ranges. These findings motivate a practical design principle: Use available physical knowledge to cover plausible model errors while preserving constraints that complement the data.

AI USE STATEMENT

Generative AI tools were used as assistive tools for writing and editing, mathematical proofreading and consistency checking, and code development and debugging. Any AI-assisted outputs incorporated into this work were reviewed and validated by the authors as appropriate. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Sec. 5.1 summarizes the setup; Apps. D.10, D.11, and F.7 document profiling and computation; App. E, App. F, and App. G provide proofs, full protocols, and additional results, respectively.

REFERENCES

James B Bartoo and Prem S Puri. On optimal asymptotic tests of composite statistical hypotheses. *The Annals of Mathematical Statistics*, 38(6):1845–1852, 1967.

Amir Beck and Aharon Ben-Tal. Duality in robust optimization: Primal worst equals dual best. *Oper. Res. Lett.*, 37(1):1–6, 2009.

Jenný Brynjarsdóttir and Anthony O'Hagan. Learning about physical parameters: The importance of model discrepancy. *Inverse problems*, 30(11):114007, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. 2018.

Suchuan Dong and Yiran Wang. A method for computing inverse parametric PDE problems with random-weight neural networks. *J. Comput. Phys.*, 489:112263, 2023.

Karthik Duraisamy, Gianluca Iaccarino, and Heng Xiao. Turbulence modeling in the age of data. *Annual review of fluid mechanics*, 51(1):357–377, 2019.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *ICML*, volume 202, pp. 10835–10866, 2023.

Roger G Ghanem and Pol D Spanos. *Stochastic finite elements: a spectral approach*. 2003.

Gene Golub and Victor Pereyra. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19(2):R1–R26, 2003.

Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973.

Gene H Golub and Charles F Van Loan. An analysis of the total least squares problem. *SIAM journal on numerical analysis*, 17(6):883–893, 1980.

Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors. *Statistical science*, 14(4):382–417, 1999.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural Operator: Learning Maps Between Function Spaces With Applications to PDEs. *J. Mach. Learn. Res.*, 24:89:1–89:97, 2023.

Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *NeurIPS*, pp. 26548–26560, 2021.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. In *ICLR*, 2021.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.*, 3(3):218–229, 2021.

Levi D. McClenny and Ulisses M. Braga-Neto. Self-adaptive physics-informed neural networks. *J. Comput. Phys.*, 474:111722, 2023.

Mario Milanese and Antonio Vicino. Optimal estimation theory for dynamic systems with set membership uncertainty: An overview. *Autom.*, 27(6):997–1009, 1991.

Susan A Murphy and Aad W Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000.

Elizabeth Newman, Lars Ruthotto, Joseph L. Hart, and Bart G. van Bloemen Waanders. Train Like a (Var)Pro: Efficient Training of Neural Networks with Variable Projection. *SIAM J. Math. Data Sci.*, 3(4):1041–1066, 2021.

Matthew Plumlee. Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285, 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. *J. Comput. Phys.*, 378:686–707, 2019.

Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science*, 367(6481):1026–1030, 2020.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in Training PINNs: A Loss Landscape Perspective. In *ICML*, volume 235, pp. 42159–42191, 2024.

Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinform.*, 25(15):1923–1929, 2009.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An Extensive Benchmark for Scientific Machine Learning. In *NeurIPS*, 2022.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. In *NeurIPS*, 2020.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, pp. 23–30, 2017.

Lloyd N Trefethen. *Spectral methods in MATLAB*. 2000.

Yannick Tr'emolet. Accounting for an imperfect model in 4D-Var. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 132(621):2483–2504, 2006.

Rui Tuo and CF Jeff Wu. Efficient calibration for imperfect computer models. 2015.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and Mitigating Gradient Flow Pathologies in Physics-Informed Neural Networks. *SIAM J. Sci. Comput.*, 43(5):A3055–A3081, 2021.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *J. Comput. Phys.*, 449:110768, 2022.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the econometric society*, pp. 1–25, 1982.

Jin-Long Wu, Jian-Xun Wang, and Heng Xiao. A Bayesian calibration–prediction method for reducing model-form uncertainties with application in RANS simulations. *Flow, Turbulence and Combustion*, 97(3):761–786, 2016.

Zongren Zou, Xuhui Meng, and George Em Karniadakis. Correcting model misspecification in physics-informed neural networks (PINNs). *J. Comput. Phys.*, 505:112918, 2024.

APPENDIX TABLE OF CONTENTS

A Notation **14**

B Related Work **15**

C Extended Analysis of Misspecified Physics Bias **16**

D Extended Theory of Compatibility Training **17**

E Proofs of Formal Statements **32**

F Experimental Protocol and Reproducibility **45**

G Additional Results **54**

H Limitations **67**

Table 5: List of notation. A star marks a true, data-generating quantity and a hat marks the nominal quantity that is encoded in the training residual.

Symbol	Meaning
$\mathcal{H}, \mathcal{Y}, \mathcal{Z}$	Hilbert spaces of predictions, residuals, observations
$\langle\cdot,\cdot\rangle, \|\cdot\|, (\cdot)^*$	Inner product, norm, Hilbert adjoint
$u, u^\star$	Prediction; true solution
f_θ, θ	Predictor model; trainable parameters
$\mathcal{N}, \mathcal{V}$	Local basin around $u^\star$; admissible predictor set
O, y, ξ	Observation operator; observations; observation noise
$\mu, \mu^\star, \hat{\mu}$	Physics parameter; true value; nominal value
Δ_μ, Δ	Mismatches $\hat{\mu} - \mu^\star$ and $\mu^\star - \hat{\mu}$
$p, \mathcal{U}$	Parameter dimension; compact uncertainty set
$\mathrm{dist}(\mu^\star, \mathcal{U})$	Distance from the true physics to the uncertainty set
$R_\mu(u)$	Physics residual under parameter μ
$r_\mu(u; Z), Z$	Pointwise residual; collocation point
$\eta_{\mathcal{U}}$	Minimum residual norm at $u^\star$ over the declared family
e	Residual discrepancy at the truth before nuisance adjustment
$\Sigma^{1/2}$	Square root of the parameter covariance Σ
$\mathcal{L}_{\mathrm{data}}, \lambda$	Data loss; physics weight
$P_{\hat{\mu}}$	Nominal physics penalty $\|R_{\hat{\mu}}(u)\|^2$
$\mathcal{C}_{\mathcal{U}}$	Compatibility loss $\min_{\mu\in\mathcal{U}} \|R_\mu(u)\|^2$
$\mathcal{C}_{\mathrm{lin}}$	Affine local profile $\min_v \|e + Ah + Jv\|^2$
$\mathcal{B}_{\hat{\mu}}$	Tangent-projected residual loss
$\mathcal{J}_{\mathrm{nom}}, \mathcal{J}_{\mathrm{comp}}, \mathcal{J}_{\tan}$	Nominal, compatibility, tangent-projected objectives
$u_{\lambda,\hat{\mu}}, u_\lambda$	Minimizers of nominal and compatibility objectives
$\mu_{\mathrm{prof}}, \mu_{\mathrm{loc}}, \bar{\mu}$	Profiled nuisance parameter: implemented, local, unconstrained
A	State derivative $D_u R_\mu(u^\star)$ at the stated evaluation point
J	Nuisance Jacobian $\partial_\mu R_\mu(u^\star)$
$\mathcal{T}_{\hat{\mu}}(u), \Pi_{\mathcal{T}}$	Nuisance tangent space $\mathrm{Range}(J)$; the tangent projector
Q	Projector $I - \Pi_{\mathrm{Range}(J)}$ off the nuisance tangent
m, M	Strong-convexity and gradient-Lipschitz constants of $\mathcal{L}_{\mathrm{data}}$
K	Lipschitz constant of $\nabla P_{\hat{\mu}}$
B, L_μ, H_μ	Bounds on $D_u R_\mu$, $\partial_\mu R_\mu$, $\partial^2_{\mu\mu} R_\mu$
B_T, σ	Tangent-profile Lipschitz constant; lower bound for J
g	Nominal physics gradient size $\|\nabla_u P_{\hat{\mu}}(u^\star)\|$
ν, c_R	Adjoint and residual nondegeneracy constants
c_{id}, γ	Identifiability and joint-observability constants
ε	Bound on the data gradient at $u^\star$
$n_c, \ell_{u,\mu}$	Collocation sample size; pointwise squared residual
$\mathcal{K}, \mathfrak{N}$	Joint class $\mathcal{V} \times \mathcal{U}$; the covering number of that class
$\delta, \rho, n_{\mathrm{obs}}$	Mismatch ratio; uncertainty radius; observation count
RelL2, GDS	Prediction-error and degradation metrics (Sec. 5.1)
$\widehat{\mathrm{Res}}_\mu$	Empirical residual MSE reported in the tables
$\nu^\star, \hat{\nu}$	True and nominal Burgers viscosity
D, α, β	Diffusivity, reaction rate, reaction-law exponent
κ, a	Combined viscosity $\zeta + 4\eta/3$ and Darcy permeability field

A NOTATION

Table 5 collects the symbols used in the main text and in the proofs, running from the underlying spaces and predictors to the experimental parameters and metrics.

B Related Work

We focus on the connections to physics-informed learning, model discrepancy, and profiling that clarify how equation uncertainty affects the training loss and predictor recovery.

Physics-informed machine learning with a fixed residual. PINNs introduced the use of PDE residuals as soft constraints for forward and inverse problems (Raissi et al., 2019), with inverse applications that reconstruct hidden velocity and pressure fields from partial observations of a known governing model (Raissi et al., 2020); the broader physics-informed machine learning program is surveyed by Karniadakis et al. (2021), and neural operators extend the idea to operator learning (Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023; Li et al., 2024). Most methods in this line assume that the encoded residual is correct. Work on PINN optimization pathologies, loss imbalance, ill-conditioned residual loss landscapes, and adaptive residual weighting studies failures that can occur even with correct physics (Wang et al., 2021; Krishnapriyan et al., 2021; Wang et al., 2022; McClenny & Braga-Neto, 2023; Rathore et al., 2024); those pathologies are properties of the optimization path and are addressed by better conditioning or reweighting, whereas the bias of Theorem 1 is a property of the stationary points themselves: The bias survives exact optimization, and rescaling a residual that points in the wrong direction shifts rather than removes the bias floor.

Misspecified models, calibration, and discrepancy. Optimizing a misspecified criterion can yield a biased pseudo-true optimizer instead of the truth, a classical statistical observation (White, 1982). The same effect appears whenever training pushes harder on a proxy that only approximates the target, as in reward-model overoptimization for preference-based fine-tuning (Gao et al., 2023). Theorem 1 is the residual-geometry version of that statement. Misspecification is likewise the central concern of Bayesian calibration of computer models, where a discrepancy process is added to the simulator so that calibration parameters are not forced to absorb structural error (Kennedy & O'Hagan, 2001); omitting that term is known to yield biased and overconfident parameter estimates (Brynjarsdóttir & O'Hagan, 2014). Including a discrepancy term has a dual cost: The calibration parameter becomes only weakly identified, which motivates calibration criteria defined by a projection onto the model family rather than by a true parameter value (Tuo & Wu, 2015) and discrepancy priors constrained to be orthogonal to the derivative of the simulator with respect to the calibration parameters (Plumlee, 2017). That orthogonality is the statistical counterpart of the tangent projection of Sec. D.9, applied to a prior rather than to a training residual. Projected calibration replaces the target of estimation but still reports a parameter as the object of interest, and projected-calibration guarantees concern that parameter; here parameter accuracy is a diagnostic, while the principal guarantees concern the predictor.

Learned model-discrepancy corrections. Within physics-informed learning, Zou et al. (2024) address misspecified PINN physics by jointly learning a solution and a discrepancy network, with Bayesian and ensemble variants for uncertainty quantification. The experiments with learned discrepancy corrections include reaction–diffusion and sparse or noisy observations, making discrepancy learning a close comparison in problem setting. Both approaches specify a discrepancy class. Compatibility training profiles over a declared family to produce a prediction loss; the selected correction is used to evaluate that loss and can be inspected as a diagnostic. For matched deterministic objectives, joint correction fitting and profiling share globally optimal predictors, as explained in Sec. D.3. Our analysis characterizes the prediction consequences of this freedom: which mismatch directions are absorbed, which physics constraints survive, and when those constraints complement incomplete observations. These questions distinguish predictor recovery from identifying a unique correction and connect model-discrepancy design to the physics constraints that remain after profiling.

Profiling and nuisance elimination. The profiled objective is a partial minimization of the joint inverse objective (Proposition 2), following variable projection in separable nonlinear least squares (Golub & Pereyra, 1973; 2003). The profiled objective also connects to profile likelihood in semiparametric statistics, where the curvature of the profiled criterion is the efficient information obtained after eliminating the nuisance parameter (Murphy & Van der Vaart, 2000); Theorem 4 is the residual-geometry analogue of that Schur-complement statement. Variable projection has also been used inside neural-network training, where the linear output-layer weights are profiled out of the

data-fitting loss to accelerate and stabilize optimization (Newman et al., 2021); there the profiled variable is a network parameter in the data term and the aim is optimization efficiency, whereas here the profiled variable is a physical parameter in the residual term. More directly, Dong & Wang (2023) eliminate inverse physical parameters with a variable-projection algorithm for neural inverse PDE solvers. Our contribution concerns the prediction consequences of this elimination under a declared uncertainty family: nominal misspecification bias, retained physics curvature, and the effects of nuisance granularity and finite collocation. In dynamical models, the profile is also used diagnostically to separate structural from practical non-identifiability of a calibration parameter (Raue et al., 2009). Our Proposition 3 likewise requires a parameter-identifiability condition for the parameter-error bound, separately from predictor recovery. Treating the physics parameter as a nuisance variable connects the tangent form to the orthogonalized scores introduced by Bartoo & Puri (1967) and to debiased machine learning (Chernozhukov et al., 2018): Sensitivity to the nuisance physics direction is removed from the residual before the residual is used as a training signal. Minimizing over admissible perturbations of the operator rather than of the residual alone is the classical total-least-squares view of errors in variables (Golub & Van Loan, 1980), which our objective transposes from parameter estimation to predictor training. Profiling is therefore classical; our contribution is to use profiling and tangent orthogonalization as a residual-geometry principle for misspecified physics-informed machine learning.

Compatibility and uncertainty-average losses. The compatibility loss takes the smallest residual over an admissible equation family, making it a best-case objective. Best-case and worst-case formulations are linked by the duality between primal worst and dual best (Beck & Ben-Tal, 2009). Taking the best rather than the worst member of an admissible family is the logic of set-membership, or bounded-error, estimation, where a candidate is acceptable when some parameter in the admissible set explains the observations (Milanese & Vicino, 1991); compatibility training uses the minimum residual over that set as the predictor's physics loss. A third option is to average over the admissible family rather than to take the best or worst member, which is the logic of Bayesian model averaging (Hoeting et al., 1999) and of domain randomization in sim-to-real transfer, where a policy is trained across a randomized range of simulator parameters (Tobin et al., 2017), and which our uncertainty-average baseline implements at the level of the residual. All three options depend on how the admissible set is built and, for field-valued uncertainty, on how a random coefficient field is reduced to finitely many modes, as in spectral stochastic finite-element parameterizations (Ghanem & Spanos, 2003); our radius sweeps probe the same sensitivity empirically.

C Extended Analysis of Misspecified Physics Bias

Let the true solution $u^\star$ satisfy $R_{\mu^\star}(u^\star) = 0$, but suppose training uses a nominal parameter $\hat{\mu} = \mu^\star + \Delta_\mu$. The nominal objective is

$$\mathcal{J}_{\mathrm{nom}}(u) \coloneqq \mathcal{L}_{\mathrm{data}}(u) + \lambda P_{\hat{\mu}}(u), \qquad P_{\hat{\mu}}(u) \coloneqq \|R_{\hat{\mu}}(u)\|^2.$$

When $(u, \mu) \mapsto R_\mu(u)$ is jointly smooth near $(u^\star, \mu^\star)$, the nominal residual and the gradient of the squared residual norm with respect to the predictor admit first-order expansions in the parameter mismatch. The following lemma gives the precise hypotheses for the expansion used in Sec. 2.

Lemma 1 (First-order expansion of the nominal gradient)**.** *Let $R_{\mu^\star}(u^\star) = 0$ and let $N \subseteq \mathbb{R}^p$ be a convex neighborhood of $\mu^\star$. Assume that $\mu \mapsto R_\mu(u^\star)$ is twice continuously differentiable on N, that $u \mapsto R_\mu(u)$ is Fréchet differentiable at $u^\star$ for every $\mu \in N$, and that $\mu \mapsto D_u R_\mu(u^\star)$ is locally Lipschitz at $\mu^\star$. Then, as $\hat{\mu} \to \mu^\star$ within N, writing $\Delta_\mu \coloneqq \hat{\mu} - \mu^\star$,*

$$R_{\hat{\mu}}(u^\star) = \partial_\mu R_{\mu^\star}(u^\star)\Delta_\mu + O(\|\Delta_\mu\|^2),$$
$$\nabla_u P_{\hat{\mu}}(u^\star) = 2 D_u R_{\mu^\star}(u^\star)^* \partial_\mu R_{\mu^\star}(u^\star)\Delta_\mu + O(\|\Delta_\mu\|^2).$$

The leading gradient is generally nonzero, so the true solution need not be stationary for the nominal physics term. The proof of Lemma 1 appears in App. E. The next result transfers the gradient perturbation to a predictor-error bound.

Proposition 1 (Nominal residual bias)**.** *Let $\mathcal{N} \subseteq \mathcal{H}$ be an open convex neighborhood of $u^\star$, and fix $\lambda \geq 0$. Assume that $\mathcal{L}_{\mathrm{data}}$ is Fréchet differentiable and m-strongly convex on $\mathcal{N}$ with $\nabla \mathcal{L}_{\mathrm{data}}(u^\star) =$*

0, *that* $P_{\hat{\mu}}$ *is Fréchet differentiable on* $\mathcal{N}$ *with* K*-Lipschitz gradient, and that a local minimizer* $u_{\lambda,\hat{\mu}} \in \mathcal{N}$ *of* $\mathcal{J}_{\mathrm{nom}}$ *satisfies* $\lambda K < m$*. Then*

$$\|u_{\lambda,\hat{\mu}} - u^\star\| \leq \frac{\lambda}{m - \lambda K} \|\nabla P_{\hat{\mu}}(u^\star)\|.$$

The proposition is local and isolates the failure channel. The resulting bound worsens with λ, showing how a wrong residual can amplify bias as the physics weight increases.

Proposition 1 alone does not prove that wrong physics is harmful: An upper bound is consistent with no bias. Theorem 1, stated in Sec. 2, closes this gap with a lower bound. The lower-bound theorem requires no convexity of the data term, only smoothness, and applies to every interior stationary point of the nominal objective in the basin.

Corollary 1 (First-order nominal bias under parameter mismatch)**.** *Under the hypotheses of Theorem 1, assume in addition that, writing* $\Delta_\mu := \hat{\mu} - \mu^\star$*,* $R_{\hat{\mu}}$ *is Fréchet differentiable at* $u^\star$*, that the linearized operator* $A := D_u R_{\hat{\mu}}(u^\star)$ *satisfies* $\|A^* y\| \geq \nu \|y\|$ *for all* $y \in \mathcal{Y}$ *with* $\nu > 0$*, and that* $\|R_{\hat{\mu}}(u^\star)\| \geq c_R \|\Delta_\mu\|$ *for some* $c_R > 0$*. Then* $g \geq 2\nu c_R \|\Delta_\mu\|$*, and every stationary point* $u_{\lambda,\hat{\mu}} \in \mathcal{N}$ *satisfies*

$$\|u_{\lambda,\hat{\mu}} - u^\star\| \geq \frac{2\nu c_R \lambda}{M + \lambda K} \|\hat{\mu} - \mu^\star\|.$$

Together with Proposition 1, the two results sandwich the nominal-training bias between $\lambda g/(M + \lambda K)$ and $\lambda g/(m - \lambda K)$ whenever the common hypotheses of the two results hold. Under the upper smoothness assumptions of Lemma 1 and the nondegeneracy hypotheses of Corollary 1, one additionally has $g = \Theta(\|\hat{\mu} - \mu^\star\|)$ when those constants are uniform over the mismatch neighborhood, so the bias is of exactly first order in the model error.

The lower-bound expression increases monotonically in λ and approaches the asymptote g/K when $K > 0$. Under the stated nondegeneracy conditions, no stationary point in the basin can be unbiased at a positive physics weight: Tuning trades the size of the guaranteed bias against the strength of the physics signal, while $\lambda = 0$ removes both. The monotonicity also has a consequence for constraint enforcement: A large physics weight is the penalty counterpart of imposing the equation exactly, so the strength of enforcement that is most useful when the encoded equation is correct is exactly what maximizes the guaranteed bias when that equation is wrong. This monotone bias-floor mechanism is consistent with the anti-informative λ-sweep behavior that we observe in Table 1.

The nondegeneracy conditions are explicit and checkable. The uniform adjoint lower bound of Corollary 1 holds with $\nu = 1/\|A^{-1}\|$ whenever the linearized nominal operator is boundedly invertible, which means that the linearized equation is well posed. The residual lower-bound assumption is the same kind of parameter-identifiability condition at the true solution as used for the profile-stability guarantee of Proposition 3 below. The degenerate case $g = 0$ is equally informative: That case means the nominal residual is orthogonal to all first-order predictor-induced residual directions, so the squared nominal residual has zero predictor gradient at $u^\star$.

Complete proofs of Proposition 1 and all subsequent formal statements, together with the auxiliary lemmas that are stated only there, are given in App. E.

D Extended Theory of Compatibility Training

D.1 Global Profile

Let $\mathcal{U}$ be a compact uncertainty set. We define

$$\mathcal{C}_{\mathcal{U}}(u) := \min_{\mu \in \mathcal{U}} \|R_\mu(u)\|^2.$$

This compatibility loss measures how far a prediction is from being explainable by some member of the declared family rather than by one nominal member. The inner minimum makes $\mathcal{C}_{\mathcal{U}}$ the optimistic, or best-case, counterpart of the usual worst-case robust penalty. A worst-case residual can be poorly matched to a setting in which a single unknown physical system generated the data, as we observe in the experiments of Sec. 5.

Under uncertainty-set containment, the profile can select $\mu^\star$, so $\mathcal{C}_{\mathcal{U}}(u^\star) = 0$; whenever the envelope is differentiable, this zero minimum has vanishing gradient. If $\mu^\star \notin \mathcal{U}$, both the remaining residual incompatibility and the active-branch gradient bounds are controlled by the distance from the true physics to the uncertainty set rather than the full nominal error. Parameter-set exclusion alone does not imply a nonzero predictor bias. In particular, when $\mu \mapsto R_\mu(u^\star)$ is Lipschitz and the linearized residuals are uniformly bounded over $\mathcal{U}$, both the profiled value $\mathcal{C}_{\mathcal{U}}(u^\star)$ and the physics gradient of every active profile branch at $u^\star$ are controlled by $\operatorname{dist}(\mu^\star, \mathcal{U})$. Lemma 5 in App. E states these upper and gradient bounds.

When the envelope is differentiable, the gradient bound applies to $\nabla\mathcal{C}_{\mathcal{U}}(u^\star)$ itself, and Sec. D.6 verifies that bound at a regular profile. When the inner minimizer is not unique, such a value function is in general only locally Lipschitz, and the corresponding first-order object is a generalized gradient.

The bounds so far concern the residual and the residual gradient at the true solution, whereas the object of interest is a minimizer of the full objective $\mathcal{J}_{\mathrm{comp}}$. The following theorem bounds every compatibility-trained local minimizer by the distance from the true physics to the uncertainty set, with perturbations of the data gradient entering additively.

Theorem 2 (Local recovery under compatibility training)**.** *Let $\mathcal{U}$ be compact, let $\mathcal{N}$ be an open convex neighborhood of $u^\star$, and let $\lambda \geq 0$. Suppose $\mathcal{L}_{\mathrm{data}}$ is m-strongly convex on $\mathcal{N}$ and $\|\nabla\mathcal{L}_{\mathrm{data}}(u^\star)\| \leq \varepsilon$. Suppose also that $\sup_{u\in\mathcal{N},\mu\in\mathcal{U}} \|D_u R_\mu(u)\| \leq B$, that $R_\mu(u)$ is continuous in μ, and that $R_{\mu^\star}(u^\star) = 0$ while $R_\mu(u^\star)$ is L_μ-Lipschitz in μ. If $2B^2\lambda < m$, every local minimizer $u_\lambda \in \mathcal{N}$ of $\mathcal{J}_{\mathrm{comp}}$ satisfies*

$$\|u_\lambda - u^\star\| \leq \frac{\varepsilon + 2BL_\mu\lambda \operatorname{dist}(\mu^\star, \mathcal{U})}{m - 2B^2\lambda}. \tag{4}$$

Thus containment and a vanishing data gradient give exact local recovery, whereas exclusion of the truth degrades continuously with $\operatorname{dist}(\mu^\star, \mathcal{U})$.

An approximate family need not contain any equation that vanishes exactly at the true solution. The same argument controls prediction through the remaining residual, without requiring a true parameter in the chosen parameterization.

Corollary 2 (Recovery for an approximate residual family)**.** *Let $\mathcal{U} \subset \mathbb{R}^p$ be nonempty and compact, let $\mathcal{N} \subseteq \mathcal{H}$ be an open convex neighborhood of $u^\star$, and let $\lambda \geq 0$. Suppose $\mathcal{L}_{\mathrm{data}}$ is Fréchet differentiable and m-strongly convex on $\mathcal{N}$, with $\|\nabla\mathcal{L}_{\mathrm{data}}(u^\star)\| \leq \varepsilon$. For every $\mu \in \mathcal{U}$, suppose $u \mapsto R_\mu(u)$ is Fréchet differentiable on $\mathcal{N}$, with $\sup_{u\in\mathcal{N},\mu\in\mathcal{U}} \|D_u R_\mu(u)\| \leq B$, and suppose $\mu \mapsto R_\mu(u)$ is continuous for every $u \in \mathcal{N}$. Define*

$$\eta_{\mathcal{U}} \coloneqq \min_{\mu\in\mathcal{U}} \|R_\mu(u^\star)\|.$$

If $2\lambda B^2 < m$, every local minimizer $u_\lambda \in \mathcal{N}$ of $\mathcal{J}_{\mathrm{comp}}$ satisfies

$$\|u_\lambda - u^\star\| \leq \frac{\varepsilon + 2\lambda B\eta_{\mathcal{U}}}{m - 2\lambda B^2}. \tag{5}$$

The proof in App. E.6 stops the recovery argument before converting the residual into a parameter distance. Under the true-model and parameter-Lipschitz assumptions of Theorem 2, $\eta_{\mathcal{U}} \leq L_\mu \operatorname{dist}(\mu^\star, \mathcal{U})$, recovering Eq. 4. The quantity $\eta_{\mathcal{U}}$ measures family approximation in the analysis; $\eta_{\mathcal{U}}$ is not needed to construct the family or evaluate the training loss. This result concerns exact profiles and a locally strongly convex data loss. Joint data–physics identifiability under incomplete observations is treated separately in Sec. D.6.

The stationarity argument in Theorem 2 requires $2B^2\lambda < m$. A comparison of objective values removes this restriction for minimizers over the basin, at the price of a weaker rate. Let $\bar{\mathcal{V}}$ be a closed ball centered at $u^\star$ within that basin, and let u_λ minimize $\mathcal{J}_{\mathrm{comp}}$ over $\bar{\mathcal{V}}$. Then

$$\|u_\lambda - u^\star\| \leq \frac{2\varepsilon}{m} + L_\mu\sqrt{\frac{2\lambda}{m}} \operatorname{dist}(\mu^\star, \mathcal{U}). \tag{6}$$

This bound holds for every $\lambda \geq 0$ and uses neither the uniform residual-derivative bound nor any first-order condition. The bound depends on λ through $\sqrt{\lambda}$ rather than λ, reflecting the loss of first-order information at the minimizer.

Theorem 2 is the predictor-level counterpart of Proposition 1: Both bounds contain a physics weight working against a conditioning margin, but the residual scale changes. The nominal bound depends on the full physics gradient $\|\nabla P_{\hat{\mu}}(u^\star)\|$, whereas the compatibility recovery bound depends on the distance from the truth to the uncertainty set. Under the nondegeneracy hypotheses of Corollary 1, the matching lower bound shows that, for every fixed positive physics weight, every stationary point in the basin obeys a first-order-in-mismatch bias floor. For compatibility training, the scale is $2BL_\mu \operatorname{dist}(\mu^\star, \mathcal{U})$, which vanishes as soon as the uncertainty set contains the truth and remains controlled as the set is shrunk past the truth. This behavior is probed by the radius sweeps of Tables 28 and 29.

The perturbation level ε makes the statement meaningful for finite, noisy data, where the empirical data gradient at $u^\star$ need not vanish: The numerator contains no explicit $\lambda\varepsilon$ cross-term. The overall noise sensitivity can nevertheless worsen as λ approaches the stability threshold because the denominator $m - 2B^2\lambda$ shrinks. This distinction is consistent with the sparse and noisy phase grids below. The value-comparison bound in Eq. 6 complements Eq. 4: The value-comparison bound requires no upper bound on λ for a minimizer over a closed basin restriction, at the price of a weaker rate.

Under uncertainty-set containment and $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$, Theorem 2 forces every local compatibility minimizer in the basin to equal $u^\star$ whenever $2B^2\lambda < m$. Combining this specialization with Theorem 1 yields a strict separation between the two residual geometries in the same basin: Under the stated conditions, profiling is not merely no worse than obedience to the nominal equation; profiling is provably better whenever the nominal physics gradient at the truth is nondegenerate.

Corollary 3 (Strict separation of nominal and compatibility training)**.** *Let all assumptions of Theorem 2 hold with $\mu^\star \in \mathcal{U}$ and $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$, and let all hypotheses of Theorem 1 hold on the same neighborhood $\mathcal{N}$. Then for every $\lambda > 0$ satisfying $2B^2\lambda < m$, the set of local minimizers of $\mathcal{J}_{\mathrm{comp}}$ in $\mathcal{N}$ is exactly $\{u^\star\}$, whereas every local minimizer $u \in \mathcal{N}$ of $\mathcal{J}_{\mathrm{nom}}$ satisfies $\|u - u^\star\| \geq \lambda g/(M + \lambda K) > 0$.*

The corollary makes precise what unbiasedness of a physics loss means under model uncertainty: With the same data term, the same basin, and the same physics weight, obeying one wrong equation biases every reachable solution, while remaining compatible with a family that contains the truth does not.

D.2 Relation to Inverse PINNs and Variable Projection

We next clarify how minimizing the residual over the allowed physics parameters before each predictor update relates to jointly learning the predictor and those parameters. The following proposition shows that, with the same family and loss, optimizing out the physics parameters preserves the globally optimal predictions.

Proposition 2 (Equivalence to joint inverse optimization)**.** *For the same predictor class, data loss, residual, weight $\lambda > 0$, and nuisance set $\mathcal{U}$, let $F(\theta, \mu) \coloneqq \mathcal{L}_{\mathrm{data}}(f_\theta) + \lambda\|R_\mu(f_\theta)\|^2$. If every inner minimum is attained, then*

$$\inf_\theta \mathcal{J}_{\mathrm{comp}}(f_\theta) = \inf_{\theta, \mu\in\mathcal{U}} F(\theta, \mu).$$

Whenever global minima are attained, the globally optimal predictor sets coincide.

Proposition 2 states the exact relationship to joint inverse optimization. Let Θ denote the predictor parameter set and write

$$M(\theta) \coloneqq \arg\min_{\mu\in\mathcal{U}} F(\theta, \mu), \qquad \mathcal{J}_{\mathrm{comp}}(f_\theta) = \min_{\mu\in\mathcal{U}} F(\theta, \mu).$$

Then the full minimizer relation is

$$\arg\min_{\theta\in\Theta, \mu\in\mathcal{U}} F(\theta, \mu) = \{(\theta, \mu) : \theta \in \arg\min_{\vartheta\in\Theta} \mathcal{J}_{\mathrm{comp}}(f_\vartheta) \text{ and } \mu \in M(\theta)\}. \tag{7}$$

Compactness of $\mathcal{U}$ and continuity in μ ensure inner attainment. Neither convexity, parameter identifiability, nor true-model containment is needed for the equivalence; the infimum identity also holds without attainment if inner minima are replaced by infima. The proof is in App. E.5. Matching the formulations requires the same residual norm, data term, weight, nuisance constraints and sharing pattern; any additional nuisance regularizer must also be included in both problems.

This nuisance elimination is classical variable projection in nuisance-affine least squares and partial minimization more generally (Golub & Pereyra, 1973; 2003; Dong & Wang, 2023). Treating the selected parameter as a nuisance specifies the scientific target and the training procedure; the nuisance interpretation does not change the matched global predictor set. Predictor guarantees for a global compatibility minimizer apply equally to the predictor of a matched joint global minimizer on the same domain. Profiling evaluates the inner optimum for the current predictor, whereas joint gradient training evolves a persistent parameter alongside the predictor. The finite optimization paths of profiling and joint gradient training can differ.

The uncertainty model specifies the unit of profiling. For finitely many instance losses $g_i(\theta, \mu_i)$ with independent attained inner minima,

$$\sum_i \min_{\mu_i \in \mathcal{U}_i} g_i(\theta, \mu_i) = \min_{(\mu_i) \in \prod_i \mathcal{U}_i} \sum_i g_i(\theta, \mu_i). \tag{8}$$

Thus sample-specific and field-valued profiles are also reduced joint objectives when the joint formulation has the same nuisance freedom. Requiring a common parameter instead restricts that freedom and generally changes the model. A trajectory may share one coefficient across all collocation points on that trajectory, while an operator dataset may require a different field correction for each input function. The field-valued Darcy comparison tests this granularity choice, rather than superiority over a matched field-inverse solver. Enlarging the nuisance family can absorb more mismatch but can also remove useful physics information, as quantified in Corollary 7.

Predictor identifiability remains necessary for predictor recovery in either formulation; eliminating a parameter does not create missing information. Theorem 2 permits multiple nuisance minimizers, whereas the smooth Hessian theory assumes a regular unique profile. The next result adds a parameter-identifiability condition and shows how a selected nuisance inherits accuracy from an accurate predictor. Parameter accuracy is an additional diagnostic, not a prerequisite imposed by the basic recovery bound.

Proposition 3 (Profile stability)**.** *Let the hypotheses of Theorem 2 hold with $2B^2\lambda < m$, and assume in addition the identifiability lower bound*

$$\|R_\mu(u^\star)\| \geq c_{\mathrm{id}} \|\mu - \mu^\star\| \qquad \text{for all } \mu \in \mathcal{U},$$

with $c_{\mathrm{id}} > 0$. Let $u_\lambda \in \mathcal{N}$ be any local minimizer of $\mathcal{J}_{\mathrm{comp}}$ and let $\mu_\lambda \in \arg\min_{\mu \in \mathcal{U}} \|R_\mu(u_\lambda)\|^2$ be any associated profiled parameter. Then

$$\|\mu_\lambda - \mu^\star\| \leq \frac{1}{c_{\mathrm{id}}} (L_\mu \operatorname{dist}(\mu^\star, \mathcal{U}) + 2B\|u_\lambda - u^\star\|). \tag{9}$$

Combining Eq. 9 with Theorem 2 gives the fully explicit estimate

$$\|\mu_\lambda - \mu^\star\| \leq \frac{1}{c_{\mathrm{id}}} \left(L_\mu \operatorname{dist}(\mu^\star, \mathcal{U}) + 2B \frac{\varepsilon + 2BL_\mu \lambda \operatorname{dist}(\mu^\star, \mathcal{U})}{m - 2B^2\lambda} \right). \tag{10}$$

The lower bound is a conditional parameter-identifiability property of the residual at the true solution. The identifiability lower bound separates parameter recovery from predictor recovery: Under this additional condition, nuisance error is controlled by predictor error and uncertainty-set misspecification. This interpretation also applies to a matched joint optimum by Proposition 2. The profile diagnostics of Sec. 5 illustrate the inheritance for the trained predictors; differences from jointly trained inverse rows concern the implemented training procedures and do not establish a separation between the ideal global optima of compatibility training and joint inverse training.

D.3 Relation to Structured Discrepancy Correction

For a nominal residual $R_{\hat{\mu}}$, define the structured correction $d_\mu(u) := R_\mu(u) - R_{\hat{\mu}}(u)$. Then

$$\mathcal{C}_{\mathcal{U}}(u) = \min_{\mu \in \mathcal{U}} \|R_{\hat{\mu}}(u) + d_\mu(u)\|^2. \tag{11}$$

This identity expresses the compatibility loss as the residual profile of an admissible correction family. The correction can depend on the prediction. For reaction–diffusion, $R_{D,\alpha}(u) = u_t - Du_{xx} - \alpha u(1-u)$ gives

$$d_{D,\alpha}(u) = -(D - \hat{D})u_{xx} - (\alpha - \hat{\alpha})u(1-u).$$

Thus inverse training and compatibility training adjust coefficients of the same solution-dependent correction directions, by joint updates and profiling, respectively. The field-valued Darcy family instead adjusts the coefficient through eight prescribed spatial modes.

If a deterministic correction model has the same residual family, predictor class, data loss, weight, constraints, and nuisance sharing pattern, the partial-minimization argument of Proposition 2 applies unchanged. With attained minima, jointly fitting and profiling the correction yield the same globally optimal predictors. Any correction regularizer must be included in both objectives and inside the profile. Bayesian integration over corrections is a different operation from this deterministic minimization.

The geometry in App. D.4 explains how the chosen correction family affects prediction. In the affine local model $e + Ah + Jv$, profiling retains $Qe + QAh$, where $Q = I - \Pi_{\mathrm{Range}(J)}$. The unrepresented discrepancy acts through A^*Qe, while the retained curvature is $2A^*QA$. At a regular interior compatible profile, Theorem 4 gives the corresponding exact Hessian for a nonlinear family. Corollary 7 quantifies the curvature lost when correction directions are added at fixed A, and Theorem 3 describes the statistical cost through the joint predictor–nuisance class. These results provide criteria for choosing structured discrepancy families that absorb model error while retaining constraints that complement the data.

D.4 Model Discrepancy and Recovery

The projector $Q = I - \Pi_{\mathrm{Range}(J)}$ also identifies the part of model discrepancy that parameter adjustment cannot remove. To see how this remaining discrepancy affects prediction, add a residual perturbation e to the affine local model in Eq. 2, giving $r(h, v) = e + Ah + Jv$. Here e is the residual discrepancy at the true solution ($h = 0$) before adjusting the physics parameter. Profiling the unrestricted correction v gives

$$\mathcal{C}_{\mathrm{lin}}(h) \coloneqq \min_{v} \|e + Ah + Jv\|^2 = \|Qe + QAh\|^2,$$
$$\nabla\mathcal{C}_{\mathrm{lin}}(0) = 2A^*Qe, \qquad D^2\mathcal{C}_{\mathrm{lin}} = 2A^*QA.$$

Profiling absorbs the component of e in $\mathrm{Range}(J)$ and leaves Qe. This unrepresented component contributes the predictor gradient $2A^*Qe$ at $h = 0$ and can therefore bias prediction, while QA determines which prediction changes the remaining physics constraints can detect. These identities are exact for this affine model. Theorem 5 uses these identities to bound global recovery error for residuals jointly affine in the predictor and physics parameter. The theorem assumes that the unconstrained profiles lie in $\mathcal{U}$. Proposition 5 also includes errors from computing the inner profile and optimizing the outer predictor.

Under a locally strongly convex data loss, Theorem 2 (App. D.1) bounds predictor error by data-gradient perturbation and $\mathrm{dist}(\mu^\star, \mathcal{U})$, the distance from the true physics parameter to the allowed set. When $\mu^\star \in \mathcal{U}$, the term proportional to $\mathrm{dist}(\mu^\star, \mathcal{U})$ in the error bound vanishes. Corollary 2 extends the predictor error bound to approximate families through the residual left at the true solution. Together with the nominal bias floor, Corollary 3 shows that, under the corollary's hypotheses and at the same positive physics weight, the local optimum of the compatibility objective is $u^\star$, while every local nominal optimum remains a positive distance away.

The practical role of the uncertainty family is therefore to absorb plausible model errors while retaining constraints that distinguish predictions fitting the same observations. In the local model, Qe tracks the unexplained model error and QA specifies which prediction changes remain detectable by physics. Sec. 4 uses these two roles to guide the choice of uncertainty directions and parameter bounds.

D.5 Finite Collocation and Profile Complexity

The preceding results use the population residual norm, whereas PINNs and physics-informed operators evaluate that norm on finitely many collocation points. A compatibility loss additionally selects a nuisance parameter from the same sample. Profiling after sampling could therefore overfit the nuisance parameter to those points. The next result controls this effect uniformly over the predictor and nuisance family; the joint covering number of the predictor–nuisance family makes the sampling guarantee depend explicitly on the complexity of the declared family.

Let $\mathcal{V} \subseteq \mathcal{H}$ be a compact convex predictor set, let $\mathcal{U} \subset \mathbb{R}^p$ be compact, and equip $\mathcal{K} = \mathcal{V} \times \mathcal{U}$ with the metric $d_{\mathcal{K}}((u,\mu),(v,\nu)) \coloneqq (\|u-v\|^2 + \|\mu-\nu\|^2)^{1/2}$. Let $Z_1, \dots, Z_{n_c}$ be independent collocation points with the same law as Z, and write $\ell_{u,\mu}(Z) \coloneqq \|r_\mu(u;Z)\|^2$. Define the population and empirical compatibility losses

$$\mathcal{C}(u) \coloneqq \min_{\mu \in \mathcal{U}} \mathbb{E}\ell_{u,\mu}(Z),$$

$$\widehat{\mathcal{C}}_{n_c}(u) \coloneqq \min_{\mu \in \mathcal{U}} \frac{1}{n_c} \sum_{i=1}^{n_c} \ell_{u,\mu}(Z_i).$$

Thus $\mathcal{C}$ is the population compatibility loss of Sec. 3.1, and $\widehat{\mathcal{C}}_{n_c}$ is the corresponding collocation estimate. Let $\mathfrak{N}(\mathcal{K}, d_{\mathcal{K}}, \tau)$ denote the size of a smallest τ-net of $\mathcal{K}$.

Theorem 3 (Uniform generalization of empirical compatibility profiles)**.** *Assume that $\ell_{u,\mu}$ is measurable, that $0 \le \ell_{u,\mu}(z) \le M_c$ for every (u,μ,z), and that $|\ell_{u,\mu}(z) - \ell_{v,\nu}(z)| \le L_c d_{\mathcal{K}}((u,\mu),(v,\nu))$ uniformly in z. Then for every $\tau > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{u \in \mathcal{V}} |\widehat{\mathcal{C}}_{n_c}(u) - \mathcal{C}(u)| \le 2L_c\tau + M_c \sqrt{\frac{\log(2\mathfrak{N}(\mathcal{K}, d_{\mathcal{K}}, \tau)/\delta)}{2n_c}}. \tag{12}$$

Dependence on nuisance dimension. Suppose additionally that $\mathcal{U}$ is contained in a Euclidean ball of radius $r_{\mathcal{U}} > 0$ in $\mathbb{R}^p$, and write $N_{\mathcal{V}}(s) \coloneqq \mathfrak{N}(\mathcal{V}, \|\cdot\|, s)$. A volumetric covering bound and a product of two $\tau/\sqrt{2}$-nets give

$$\mathfrak{N}(\mathcal{K}, d_{\mathcal{K}}, \tau) \le N_{\mathcal{V}}(\tau/\sqrt{2}) \left(1 + \frac{2\sqrt{2}r_{\mathcal{U}}}{\tau}\right)^p.$$

Indeed, a maximal s-separated subset of $\mathcal{U}$ is an s-net. The disjoint balls of radius $s/2$ centered at the points of the separated subset lie in a ball of radius $r_{\mathcal{U}} + s/2$, so comparing volumes bounds the subset's cardinality by $(1 + 2r_{\mathcal{U}}/s)^p$. Taking $s = \tau/\sqrt{2}$ and combining this net with a predictor net gives the displayed product bound. Consequently, under the hypotheses of Theorem 3, with probability at least $1-\delta$,

$$\sup_{u \in \mathcal{V}} |\widehat{\mathcal{C}}_{n_c}(u) - \mathcal{C}(u)| \le 2L_c\tau + M_c \sqrt{\frac{\log(2N_{\mathcal{V}}(\tau/\sqrt{2})/\delta) + p\log(1 + 2\sqrt{2}r_{\mathcal{U}}/\tau)}{2n_c}}. \tag{13}$$

This estimate in terms of nuisance dimension is a direct consequence of the existing uniform bound. For a target error greater than $2L_c\tau$, a fixed net scale and common bounds on L_c, M_c, $r_{\mathcal{U}}$, and predictor complexity make the sufficient collocation budget grow linearly with p. Comparisons across parameterizations must account for the normalization of each parameterization and any dimension dependence of these quantities. The bound need not be tight, especially for redundant parameterizations, and implies neither monotone growth of the observed profile gap with p nor a prediction-error rate in p/n_c.

Sampling and empirical scope. The theorem controls profile selection and evaluation on the same collocation batch. Uniformity over μ prevents the inner selection from escaping the guarantee. If B is a freshly sampled collocation batch and $g_B(u,\mu)$ is the empirical squared residual on batch B, then

$$\mathbb{E}_B \min_{\mu \in \mathcal{U}} g_B(u,\mu) \le \min_{\mu \in \mathcal{U}} \mathbb{E}_B g_B(u,\mu) = \mathcal{C}(u).$$

Equality need not hold: Profiling each batch and profiling the population residual are different operations. Theorem 3 controls the discrepancy between empirical and population profile values on a sampled batch, not unbiasedness of the stochastic profile gradient. Proposition 2 applies separately to each fixed empirical residual or to the population residual. The deterministic full-grid residuals used in the operator experiments perform the same algebraic operation but are outside this probabilistic statement. The reported experiments do not estimate the profile-value gap as a joint function of nuisance dimension and collocation count; the finer-grid operator check evaluates a different discretization. Sampling accuracy is also distinct from retained information: Exact population profiling can erase an unobserved prediction direction, as the example in App. D.6 shows.

D.6 Profiled Curvature and Physics-Assisted Identifiability

The preceding recovery bounds explain when compatibility training avoids model-mismatch bias for an already identifiable predictor. Those bounds rely on strong convexity of the data loss, however, and therefore leave open a complementary question, namely whether the part of the physics residual that survives profiling can identify directions that sparse data leave unresolved. The surviving residual can. Profiling removes only residual changes generated by nuisance parameters. Solution changes that point transversely to that nuisance tangent remain as positive curvature. The statements below make the retained-curvature claim precise by first establishing local smoothness of the profiled value function through an implicit-function argument.

We first work in the following regular-profile setting. Let $\mathcal{H}$ and $\mathcal{Y}$ be real Hilbert spaces, let $\mathcal{U} \subset \mathbb{R}^p$ satisfy $\mu^\star \in \operatorname{int}\mathcal{U}$, and let $u^\star \in \mathcal{H}$. Assume that $(u, \mu) \mapsto R_\mu(u)$ is jointly three times continuously Fréchet differentiable near $(u^\star, \mu^\star)$ and that $R_{\mu^\star}(u^\star) = 0$. Define

$$A := D_u R_{\mu^\star}(u^\star), \qquad J := \partial_\mu R_{\mu^\star}(u^\star).$$

Lemma 2 (Regular local nuisance profile)**.** *In the regular-profile setting above, suppose $J^*J \succeq \sigma^2 I_p$ for some $\sigma > 0$. Then there are neighborhoods $\mathcal{N}_0$ of $u^\star$ and $W \Subset \operatorname{int}\mathcal{U}$ of $\mu^\star$, and a unique twice continuously differentiable local profile $\mu_{\mathrm{loc}} : \mathcal{N}_0 \to W$ satisfying*

$$\partial_\mu\big(\tfrac{1}{2}\|R_\mu(u)\|^2\big)\big|_{\mu=\mu_{\mathrm{loc}}(u)} = 0.$$

After shrinking these neighborhoods if necessary, $\mu_{\mathrm{loc}}(u)$ is the unique minimizer of $\|R_\mu(u)\|^2$ over W for every $u \in \mathcal{N}_0$, and

$$D\mu_{\mathrm{loc}}(u^\star) = -(J^*J)^{-1}J^*A.$$

Corollary 4 (Smooth global compatibility profile)**.** *Under the hypotheses and notation of Lemma 2, assume in addition that $\mathcal{U}$ is compact, that R is jointly three times continuously Fréchet differentiable on an open set containing $\{u^\star\} \times \mathcal{U}$, and that $\mu^\star$ is the unique $\mu \in \mathcal{U}$ with $R_\mu(u^\star) = 0$. Then, on a sufficiently small neighborhood of $u^\star$, the global profile $\mu(u) := \arg\min_{\mu\in\mathcal{U}} \|R_\mu(u)\|^2$ is unique and agrees with $\mu_{\mathrm{loc}}(u)$. Consequently, $\mu(\cdot)$ and $\mathcal{C}_\mathcal{U}$ are twice continuously differentiable there.*

Theorem 4 (Exact profiled Hessian)**.** *Under the hypotheses and notation of Corollary 4, let*

$$Q := I - J(J^*J)^{-1}J^* = I - \Pi_{\mathrm{Range}(J)}.$$

Then the compatibility loss satisfies

$$\nabla\mathcal{C}_\mathcal{U}(u^\star) = 0, \qquad D^2\mathcal{C}_\mathcal{U}(u^\star)[h,k] = 2\langle QAh, QAk\rangle \quad (h,k \in \mathcal{H}).$$

Corollary 5 (Local recovery from joint data–physics identifiability)**.** *Under the hypotheses of Theorem 4, let, in addition, $\mathcal{L}_{\mathrm{data}}$ be twice continuously Fréchet differentiable near $u^\star$ with $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$. If for some $\lambda > 0$ and $\gamma > 0$,*

$$D^2\mathcal{L}_{\mathrm{data}}(u^\star)[h,h] + 2\lambda\|QAh\|^2 \geq \gamma\|h\|^2 \qquad \text{for every } h \in \mathcal{H}, \tag{14}$$

then $\mathcal{J}_{\mathrm{comp}}$ is $(\gamma/2)$-strongly convex on a sufficiently small ball about $u^\star$, and $u^\star$ is the unique minimizer of $\mathcal{J}_{\mathrm{comp}}$ on that ball.

Corollary 6 (Finite-dimensional joint observability criterion)**.** *Under the hypotheses and notation of Theorem 4, suppose that $\mathcal{H}$ is finite dimensional and that the data loss is twice continuously differentiable near $u^\star$, with $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$ and $D^2\mathcal{L}_{\mathrm{data}}(u^\star) = O^*O$ for a bounded linear observation operator O. Then, for any fixed $\lambda > 0$, condition Eq. 14 holds for some $\gamma > 0$ if and only if $\ker O \cap \ker(QA) = \{0\}$.*

The formula is a Schur complement for nuisance elimination. The projector Q discards exactly the residual tangent generated by changing μ; the projector does not discard solution-induced residual changes outside that tangent. Adding independent nuisance directions enlarges $\mathrm{Range}(J)$ and can reduce the curvature available for prediction. A physically targeted family can absorb represented discrepancy while retaining constraints on directions invisible to the observations. The projector describes the local geometry at a regular interior compatible profile; finite nonlinear adjustments and active parameter constraints are handled by the global profile and the distance-to-set bounds.

This combined data–physics curvature is also the Schur complement of the nuisance block in the joint inverse Hessian, obtained by minimizing over physics-parameter perturbations. Proposition 4 gives the corresponding local stationary-point relation and a condition-number bound when the Hessians are positive definite.

Proposition 4 (Joint and reduced local optimization geometry)**.** *Under the hypotheses of Theorem 4, suppose $\mathcal{H} = \mathbb{R}^d$ with the Euclidean inner product, $\mathcal{L}_{\mathrm{data}}$ is twice continuously differentiable near $u^\star$, and $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$. Fix $\lambda > 0$, let $H_d \coloneqq D^2\mathcal{L}_{\mathrm{data}}(u^\star)$, and define*

$$\mathcal{J}_{\mathrm{joint}}(u, \mu) \coloneqq \mathcal{L}_{\mathrm{data}}(u) + \lambda\|R_\mu(u)\|^2.$$

For the smooth profile $\mu(u)$ of Corollary 4, stationarity of u for $\mathcal{J}_{\mathrm{comp}}$ is equivalent to joint stationarity of $(u, \mu(u))$ throughout a sufficiently small neighborhood. At $(u^\star, \mu^\star)$ the joint Hessian and the Schur complement of the nuisance block are

$$H_{\mathrm{joint}} = \begin{pmatrix} H_d + 2\lambda A^*A & 2\lambda A^*J \\ 2\lambda J^*A & 2\lambda J^*J \end{pmatrix}, \tag{15}$$

$$H_{\mathrm{red}} \coloneqq D^2\mathcal{J}_{\mathrm{comp}}(u^\star) = H_d + 2\lambda A^*QA. \tag{16}$$

For every $h \in \mathbb{R}^d$,

$$\langle H_{\mathrm{red}}h, h\rangle = \min_{v\in\mathbb{R}^p} \left\langle H_{\mathrm{joint}}\begin{pmatrix} h \\ v \end{pmatrix}, \begin{pmatrix} h \\ v \end{pmatrix} \right\rangle. \tag{17}$$

In particular, $H_{\mathrm{joint}} \succ 0$ if and only if $H_{\mathrm{red}} \succ 0$. If the joint and reduced Hessians are positive definite, the spectral condition numbers of the two Hessians in the fixed Euclidean coordinates satisfy

$$\kappa_2(H_{\mathrm{red}}) \le \kappa_2(H_{\mathrm{joint}}). \tag{18}$$

The proposition connects the retained-curvature identity to the classical optimization geometry of variable projection. Eliminating nuisance updates can reduce the range of curvature scales seen by the outer problem while preserving the matched global predictor set. If $H_d = O^*O$, positive definiteness is exactly the joint observability condition of Corollary 6. The stationary-point correspondence also shows that profiling does not generally remove unfavorable stationary points on the regular minimizing branch. This geometric result is local to predictor coordinates; the result does not imply a convergence-rate ordering for adaptive moment estimation with decoupled weight decay (AdamW) or a positive-definite Hessian in an overparameterized neural representation. Active constraints, nonunique profiles, and inexact inner solves fall outside this regular interior statement. Training runs with active constraints, nonunique profiles, or inexact inner solves must be distinguished from the ideal geometry in Eq. 15.

The coercivity condition in Eq. 14 plays the role that a regularization term plays for an ill-posed inverse problem, except that the stabilizing curvature is supplied by the physics that survives profiling rather than by a generic smoothness penalty. Unlike Theorem 2, Corollary 5 does not require the data loss alone to be strongly convex or the physics weight to satisfy an upper bound. The corollary instead requires the scientifically more relevant joint observability condition in Eq. 14, whose finite-dimensional kernel criterion is isolated in Corollary 6. Together with Theorem 4, that condition explains how the physics constraints that remain after profiling can complement sparse data, providing a local mechanism consistent with the behavior in Tables 19 and 20.

The sense in which a larger family weakens physics can be made exact when one nuisance model refines another. The following corollary distinguishes the extra bias coverage obtained by admitting new nuisance directions from the identifiable curvature that those same directions remove.

Corollary 7 (Exact curvature loss under nuisance refinement)**.** *Suppose two regular compatibility families, indexed by $i \in \{1, 2\}$, satisfy the hypotheses of Theorem 4 at the same solution $u^\star$, have the same solution derivative $A : \mathcal{H} \to \mathcal{Y}$, and have full-rank nuisance Jacobians J_i with tangent spaces $\mathcal{T}_i \coloneqq \mathrm{Range}(J_i)$. Let $\mathcal{C}_i$ be the corresponding locally profiled losses and $Q_i \coloneqq I - \Pi_{\mathcal{T}_i}$. If $\mathcal{T}_1 \subseteq \mathcal{T}_2$ and $\mathcal{E} \coloneqq \mathcal{T}_2 \cap \mathcal{T}_1^\perp$, then for all $h, k \in \mathcal{H}$,*

$$D^2\mathcal{C}_1(u^\star)[h, k] - D^2\mathcal{C}_2(u^\star)[h, k] = 2\langle\Pi_{\mathcal{E}}Ah, \Pi_{\mathcal{E}}Ak\rangle. \tag{19}$$

In particular, $D^2\mathcal{C}_2(u^\star) \preceq D^2\mathcal{C}_1(u^\star)$ in positive-semidefinite order.

Corollary 8 (Quantitative preservation under nuisance refinement)**.** *In the setting of Corollary 7, let $O : \mathcal{H} \to \mathcal{Z}$ be bounded, let $\lambda > 0$, and define $q_i(h) \coloneqq \|Oh\|^2 + 2\lambda\|Q_iAh\|^2$. If $q_1(h) \ge \gamma_1\|h\|^2$ for every $h \in \mathcal{H}$ and, for some $\vartheta \in [0, 1)$,*

$$\|\Pi_{\mathcal{E}}Ah\| \le \vartheta\|Q_1Ah\| \qquad (h \in \mathcal{H}),$$

then

$$q_2(h) \ge (1 - \vartheta^2)\gamma_1\|h\|^2 \qquad (h \in \mathcal{H}).$$

Thus granularity has a two-sided mathematical effect. New nuisance directions can absorb additional model discrepancy, but the exact price is the positive-semidefinite curvature term in Eq. 19. The angle condition in Corollary 8 gives a usable design rule: Added discrepancy directions are safe for identification when those directions capture only a strict fraction of the previously retained solution-induced residual. These results concern adding tangent directions to a nuisance family; changing only the radius of a fixed interior parameterization instead changes feasibility and is governed by Theorem 2.

An exact example of retained and lost information. A simple example shows how adding a correction direction can remove a useful physics constraint. The true solution is $(x, y) = (0, 0)$, and only x is observed. The nominal residual $(x + a, y)$ contains an erroneous offset $a \neq 0$. Allowing an additive correction only in the first residual component, with a range large enough to cancel a, removes the nominal bias while preserving the residual y that constrains the unobserved coordinate. If the family also allows a correction in the second residual component, that correction can cancel errors in y within the allowed range. The observations and physics then no longer determine y uniquely.

The following jointly affine example isolates the effect of adding a nuisance direction. Let $u = (x, y) \in \mathbb{R}^2$, let $u^\star = (0, 0)$, and suppose only the first coordinate is observed, so that $\mathcal{L}_{\mathrm{data}}(x, y) = x^2/2$ and $O = (1, 0)$. For a fixed nominal discrepancy $a \neq 0$, take $R_{\mathrm{nom}}(x, y) = (x + a, y)$. Compare the two residual families

$$R_v^{(1)}(x, y) = (x + a + v, y), \qquad v \in [-\rho, \rho],$$
$$R_{v,w}^{(2)}(x, y) = (x + a + v, y + w), \qquad (v, w) \in [-\rho, \rho]^2,$$

where $\rho > |a|$. Both contain the true equation, with compatible parameters $v^\star = -a$ and $(v^\star, w^\star) = (-a, 0)$, respectively. On the open neighborhood

$$\mathcal{N} := \{(x, y) : |x + a| < \rho \text{ and } |y| < \rho\},$$

the unique profiles are $v(x, y) = -x - a$ and $(v(x, y), w(x, y)) = (-x - a, -y)$, both strictly within the corresponding compact parameter sets. Substitution gives $\mathcal{C}_1(x, y) = y^2$ and $\mathcal{C}_2(x, y) = 0$ exactly.

For every $\lambda > 0$, the nominal objective has the unique minimizer $(-2\lambda a/(1 + 2\lambda), 0)$ in $\mathcal{N}$. The structured objective $x^2/2 + \lambda\mathcal{C}_1(x, y) = x^2/2 + \lambda y^2$ has the unique minimizer $u^\star$, whereas the enlarged objective $x^2/2 + \lambda\mathcal{C}_2(x, y) = x^2/2$ is minimized by every $(0, y)$ with $|y| < \rho$. Thus a correction for the nominal discrepancy removes bias while retaining the constraint on the unobserved coordinate; adding a correction in that coordinate erases this constraint.

Writing $\mathcal{J}_i = \mathcal{L}_{\mathrm{data}} + \lambda\mathcal{C}_i$, we have $A = I_2$, $J_1 = (1, 0)^\top$, and $J_2 = I_2$, so

$$Q_1 = \mathrm{diag}(0, 1), \qquad Q_2 = 0, \qquad D^2\mathcal{J}_1 = \mathrm{diag}(1, 2\lambda), \qquad D^2\mathcal{J}_2 = \mathrm{diag}(1, 0).$$

The extra tangent direction removes exactly $2h_y^2$ of physics curvature, as in Corollary 7, and makes $\ker O \cap \ker(Q_2 A) = \mathrm{span}\{(0, 1)\}$. This loss of the constraint on the unobserved coordinate persists for exact population residuals, so increasing collocation accuracy alone cannot restore the missing constraint. In general nonlinear models, failure of the positive-curvature criterion alone does not imply nonuniqueness, since higher-order terms may still identify the solution. In this affine example, the objective is exactly flat in y, establishing actual nonuniqueness under the enlarged family.

For the narrower subclass that is jointly affine in the predictor and nuisance parameter, the same geometry admits a global result. This subclass has the form $R_\mu(u) = Au - b + J\mu$ with fixed linear operators A and J. The experimental Burgers, reaction–diffusion, and Darcy residuals are not of this form: Those residuals are affine in the nuisance parameter but generally nonlinear or bilinear in the predictor, and therefore have closed-form inner profiles but are covered by the local results above, not by the global theorem below. The jointly affine result also permits the two approximations present in training: The inner profile need only be solved to finite accuracy, and the outer objective need only be optimized to finite suboptimality.

D.7 Global Jointly Affine Recovery

Let $\mathcal{H}$, $\mathcal{Y}$, and $\mathcal{Z}$ be real Hilbert spaces, let $A : \mathcal{H} \to \mathcal{Y}$, $J : \mathbb{R}^p \to \mathcal{Y}$, and $O : \mathcal{H} \to \mathcal{Z}$ be bounded linear operators with J injective, and let $b \in \mathcal{Y}$. Define

$$R_\mu(u) \coloneqq Au - b + J\mu, \qquad Q \coloneqq I - J(J^*J)^{-1}J^*.$$

Here $A = D_u R_\mu(u)$ and $J = \partial_\mu R_\mu(u)$ are the constant solution and nuisance derivatives of Sec. D.6, so Q is the projector of Theorem 4. Let $\mathcal{V} \subseteq \mathcal{H}$ be a nonempty closed convex set, let $\mathcal{U} \subset \mathbb{R}^p$ be compact and convex, and assume that

$$\bar{\mu}(u) \coloneqq -(J^*J)^{-1}J^*(Au - b) \in \mathcal{U} \qquad \text{for every } u \in \mathcal{V}.$$

Theorem 5 (Global jointly affine recovery with sparse data)**.** *In the jointly affine setting above, the profiled parameter and compatibility loss satisfy, for every $u \in \mathcal{V}$,*

$$\{\bar{\mu}(u)\} = \arg\min_{\mu \in \mathcal{U}} \|R_\mu(u)\|^2, \qquad \mathcal{C}_{\mathcal{U}}(u) = \|Q(Au - b)\|^2.$$

Let $u^\star \in \mathcal{V}$, observe $y = Ou^\star + \xi$, fix $\lambda > 0$, and suppose the joint data–physics observability inequality

$$\|Oh\|^2 + 2\lambda\|QAh\|^2 \geq \gamma\|h\|^2 \qquad \text{for every } h \in \mathcal{H} \tag{20}$$

holds with $\gamma > 0$. Then the exact compatibility objective $F(u) \coloneqq \|Ou - y\|^2/2 + \lambda\mathcal{C}_{\mathcal{U}}(u)$ has a unique global minimizer u_{aff} on $\mathcal{V}$. For any $\mu^\star \in \mathcal{U}$, writing $e \coloneqq R_{\mu^\star}(u^\star)$, the minimizer satisfies

$$\|u_{\mathrm{aff}} - u^\star\| \leq \frac{\|O^*\xi\| + 2\lambda\|A^*Qe\|}{\gamma}. \tag{21}$$

In particular, $e = \xi = 0$ implies $u_{\mathrm{aff}} = u^\star$, even if O is not injective. If $\mathcal{H}$ is finite dimensional, then for every fixed $\lambda > 0$, Eq. 20 holds for some $\gamma_\lambda > 0$ if and only if $\ker O \cap \ker(QA) = \{0\}$. Thus this kernel condition makes the observability inequality valid for every positive physics weight, with a possibly λ-dependent coercivity constant.

Proposition 5 (Stability to inexact profiling and outer optimization)**.** *In the setting of Theorem 5, let $\widetilde{\mathcal{C}}$ be any computed inner loss with*

$$\mathcal{C}_{\mathcal{U}}(u) \leq \widetilde{\mathcal{C}}(u) \leq \mathcal{C}_{\mathcal{U}}(u) + \eta_{\mathrm{prof}} \qquad (u \in \mathcal{V}),$$

for some $\eta_{\mathrm{prof}} \geq 0$, and suppose $\widetilde{u} \in \mathcal{V}$ is an $\varepsilon_{\mathrm{opt}}$-minimizer, with $\varepsilon_{\mathrm{opt}} \geq 0$, of $\widetilde{F}(u) \coloneqq \|Ou - y\|^2/2 + \lambda\widetilde{\mathcal{C}}(u)$. Then

$$\|\widetilde{u} - u^\star\| \leq \frac{\|O^*\xi\| + 2\lambda\|A^*Qe\|}{\gamma} + \sqrt{\frac{2(\varepsilon_{\mathrm{opt}} + \lambda\eta_{\mathrm{prof}})}{\gamma}}. \tag{22}$$

The two error terms in Eq. 22 have distinct meanings. Only the out-of-family component Qe can contribute physical bias, through A^*Qe; error in $\mathrm{Range}(J)$ is absorbed by the nuisance profile. A nonzero Qe need not bias the predictor if $A^*Qe = 0$. This discrepancy decomposition is the exact jointly affine counterpart of the local model in App. D.4. Observation noise and out-of-family error perturb the exact solution linearly, whereas finite inner and outer objective errors enter through the usual square-root stability of a strongly convex objective. The affine recovery theorem and the inexact-profile proposition give a value-based benchmark for sparse and noisy observations with approximate inner solves. Applying the bound requires the jointly affine structure and the stated uniform value-error bound; the nuisance-affine but predictor-nonlinear experimental residuals do not automatically satisfy these hypotheses.

The unconstrained-profile feasibility hypothesis $\bar{\mu}(u) \in \mathcal{U}$, together with a full-rank tangent parameterization, is exactly the regime in which the box-constrained profile agrees with the unconstrained tangent profile. When the unconstrained profile falls outside $\mathcal{U}$ and clipping is active, the quadratic profile representation in Theorem 5 no longer applies. The local result of Theorem 2 remains available under the hypotheses of that theorem, with set misspecification measured by $\mathrm{dist}(\mu^\star, \mathcal{U})$ rather than by the clipping distance of $\bar{\mu}(u)$.

D.8 Computation and Approximation Overview

Local tangent projection. When the residual depends nonlinearly on the uncertain parameters, solving the full profile at every predictor update can be costly. For small parameter deviations from $\hat{\mu}$, linearizing the residual in the physics parameters replaces that nonlinear search with a linear least-squares problem:

$$J_\mu(u) \coloneqq \partial_\mu R_{\hat{\mu}}(u)\Sigma^{1/2}, \qquad \mathcal{B}_{\hat{\mu}}(u) \coloneqq \min_v \|R_{\hat{\mu}}(u) + J_\mu(u)v\|^2.$$

Here Σ specifies the declared parameter uncertainty, v indexes parameter changes through $\Sigma^{1/2}v$, and $J_\mu(u)v$ is the first-order effect of these parameter changes on the residual. The minimization removes the residual component that these linearized parameter changes can explain. The minimization allows any $v \in \mathbb{R}^p$ and does not enforce the original bounds of $\mathcal{U}$. Tangent projection removes the first-order mismatch but leaves the nonlinear Taylor remainder, the error from the terms omitted by the linearization. Theorem 6 shows that the resulting predictor bias is $O(\|\mu^\star - \hat{\mu}\|^2)$ when the theorem's assumptions hold. For large parameter deviations, the omitted nonlinear terms can make the tangent approximation inaccurate (App. D.9). Full compatibility training evaluates the original residual over the allowed family. If the allowed coefficient or spatial corrections can recover the true governing equation, then $\mathcal{C}_{\mathcal{U}}(u^\star) = 0$, and the resulting mismatch bias is removed under the corresponding recovery conditions.

Affine profiles and optimization. The inner optimization in compatibility training holds the prediction u fixed and selects the physics parameter $\mu \in \mathcal{U}$ that minimizes $\|R_\mu(u)\|^2$. When the residual is affine in a scalar parameter, $R_\mu(u) = R_0(u) + J(u)\mu$ with $J(u) \neq 0$, the exact minimizing coefficient over $[\mu_{\min}, \mu_{\max}]$ is

$$\mu_{\mathrm{prof}}(u) = \mathrm{clip}_{[\mu_{\min},\mu_{\max}]}\left(-\frac{\langle R_0(u), J(u)\rangle}{\|J(u)\|^2}\right).$$

Here $R_0(u)$ collects the residual terms independent of μ, and $J(u) = \partial_\mu R_\mu(u)$ is the residual change per unit change in that coefficient. Computing the scalar profile requires the inner product $\langle R_0(u), J(u)\rangle$ and the squared norm $\|J(u)\|^2 = \langle J(u), J(u)\rangle$. The unclipped value cancels as much of $R_0(u)$ as possible along $J(u)$; clipping keeps the coefficient within the allowed interval. For affine field corrections, the analogous calculation solves a system for the spatial basis coefficients before each predictor update. The squared norm of the remaining residual is the compatibility loss $\mathcal{C}_{\mathcal{U}}(u)$; the predictor update adds the physics penalty $\lambda\mathcal{C}_{\mathcal{U}}(u)$ to the data loss.

Finite collocation. Collocation points are the locations in space or space–time where the residual of the governing equation is evaluated during training. In practice, profiling minimizes the residual MSE over a finite sample of these points. Selecting the physics parameter on the same sample can fit those points unusually well, so the sampled minimum can underestimate the population compatibility loss. For points drawn independently from the same distribution, Theorem 3 bounds the difference between sampled and population profile values by controlling sampling error uniformly over predictors and allowed physics parameters. The covering-number term in this bound measures the size of the joint predictor–parameter family at a chosen resolution. The collocation bound shows how family complexity affects the number of collocation points sufficient to control the sampling error. With parameter normalization, radius, and the other bounds fixed, the sufficient sample count in this bound grows with the number of correction coefficients. This sampling-error bound controls how accurately the finite sample estimates the compatibility loss; the family must still preserve the physics constraints needed to identify the solution.

Further details. The protocol for constructing a family and the exact example with two coordinates appear in Apps. F.2 and D.6. The tangent approximation, affine training algorithm, and regularized or approximate profiles are detailed in Apps. D.9, D.10, and D.11, respectively. App. D.5 gives the bound for finite collocation and explains how the bound depends on nuisance dimension.

D.9 Local Tangent Projection

The first-order residual directions generated by uncertain parameters form the nuisance tangent space

$$J_\mu(u) \coloneqq \partial_\mu R_{\hat{\mu}}(u)\Sigma^{1/2}, \qquad \mathcal{T}_{\hat{\mu}}(u) \coloneqq \mathrm{Range}(J_\mu(u)).$$

Let $\Pi_{\mathcal{T}_{\hat{\mu}}(u)}$ be the orthogonal projector onto this space. The tangent-projected loss is

$$\mathcal{B}_{\hat{\mu}}(u) \coloneqq \|(I - \Pi_{\mathcal{T}_{\hat{\mu}}(u)})R_{\hat{\mu}}(u)\|^2 = \min_v \|R_{\hat{\mu}}(u) + J_\mu(u)v\|^2,$$

where the right-hand side is the profiled residual of the linearized family $R_{\hat{\mu}+\Sigma^{1/2}v}(u) \approx R_{\hat{\mu}}(u) + J_\mu(u)v$.

When the true perturbation lies in the tangent model and the projector is stable, the projected nominal residual at $u^\star$ is second order in the parameter mismatch $\Delta \coloneqq \mu^\star - \hat{\mu} = -\Delta_\mu$, while the unprojected nominal residual is first order, hence $\Theta(\|\Delta\|)$ under a lower nondegeneracy bound on the linear term. Lemma 7 in App. E states both residual bounds together with the curvature and tangent-model hypotheses that both bounds require. This second-order gap is why local tangent projection is reliable near the nominal model but can fail in the large, nonlinear mismatch regime tested below.

The projected residual also has a nuisance-orthogonal interpretation: To first order, changing the nuisance physics parameter does not change the estimating residual used to train u. App. E records the projector identity behind this statement. That identity is the residual analogue of nuisance-orthogonal estimating equations in semiparametric statistics: The method does not need the profiled parameter to be estimated accurately in order to remove the first-order direction by which misspecified physics biases the predictor.

The predictor-level argument uses a stability property of the inner tangent profile that is useful independently of the final bias bound.

Lemma 3 (Lipschitz tangent-profile residual)**.** *Let $\mathcal{N} \subseteq \mathcal{H}$ be open and convex. Let $R : \mathcal{N} \to \mathcal{Y}$ and $J : \mathcal{N} \to L(\mathbb{R}^p, \mathcal{Y})$ be Fréchet differentiable with $\|D_u R(u)\| \le B_R$ and $\|D_u J(u)\| \le B_J$ on $\mathcal{N}$, where the norm for $D_u J(u)$ is induced by the operator norm on $L(\mathbb{R}^p, \mathcal{Y})$. Assume further that $\|J(u)v\| \ge \sigma\|v\|$ for all $u \in \mathcal{N}$ and $v \in \mathbb{R}^p$, with $\sigma > 0$. Set*

$$M_R \coloneqq \sup_{u\in\mathcal{N}} \|R(u)\| < \infty, \qquad B_T \coloneqq B_R + \frac{B_J M_R}{\sigma}.$$

Then the least-squares tangent profile has the unique minimizer

$$v(u) = -(J(u)^* J(u))^{-1} J(u)^* R(u), \qquad \|v(u)\| \le \frac{M_R}{\sigma},$$

and the profile residual norm $b(u) \coloneqq \min_{v\in\mathbb{R}^p} \|R(u) + J(u)v\|$ is B_T-Lipschitz on $\mathcal{N}$.

The two residual bounds above control the signal seen by the predictor, not yet the predictor itself. The following theorem combines the two residual bounds with Lemma 3 to obtain a predictor-level guarantee of the same form as Theorem 2.

Theorem 6 (Second-order bias of tangent-projected training)**.** *Let $\mathcal{N} \subset \mathcal{H}$ be an open ball centered at $u^\star$, write $\Delta \coloneqq \mu^\star - \hat{\mu}$, and let $\lambda \ge 0$. Assume that the data loss $\mathcal{L}_{\mathrm{data}}$ is Fréchet differentiable and m-strongly convex on $\mathcal{N}$ with $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$. With $R(u) \coloneqq R_{\hat{\mu}}(u)$ and $J(u) \coloneqq \partial_\mu R_{\hat{\mu}}(u)\Sigma^{1/2}$, assume that the hypotheses of Lemma 3 hold on $\mathcal{N}$, and let B_T denote the resulting Lipschitz constant. Assume further that $R_{\mu^\star}(u^\star) = 0$ while $\mu \mapsto R_\mu(u^\star)$ is twice continuously differentiable on a convex neighborhood of the segment from $\hat{\mu}$ to $\mu^\star$ with $\|\partial^2_{\mu\mu} R_\mu(u^\star)\| \le H_\mu$ there, and the true perturbation lies in the tangent model, $\partial_\mu R_{\hat{\mu}}(u^\star)\Delta \in \mathcal{T}_{\hat{\mu}}(u^\star)$, as holds whenever Σ is nonsingular. If $2B_T^2\lambda < m$, then every local minimizer $u_\lambda \in \mathcal{N}$ of $\mathcal{J}_{\tan} \coloneqq \mathcal{L}_{\mathrm{data}} + \lambda\mathcal{B}_{\hat{\mu}}$ satisfies*

$$\|u_\lambda - u^\star\| \le \frac{\lambda B_T H_\mu \|\Delta\|^2}{m - 2B_T^2\lambda}.$$

For fixed admissible λ and locally uniform B_T, H_μ, the bound is $O(\|\Delta\|^2)$.

Together with Proposition 1, Theorem 1, Corollary 1, and Theorem 2, these results yield a bias hierarchy for the three residual geometries in the same local basin. Under the corresponding nondegeneracy hypotheses, nominal training is biased at first order in $\|\Delta\|$, with weight dependence $\lambda/(m - \lambda K)$ in the corresponding upper bound, and the residual driving that bias is generally $\Theta(\|\Delta\|)$. Tangent-projected training removes the first-order direction and retains only a second-order bias, with weight dependence $\lambda/(m - 2B_T^2\lambda)$, as long as the true perturbation is visible to the tangent model, the tangent family is nondegenerate, and the constants are locally uniform. Global compatibility training removes the bias entirely whenever $\mu^\star \in \mathcal{U}$, $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$, and the stability condition $2B^2\lambda < m$ holds.

For affine parameter dependence, $H_\mu = 0$, so Theorem 6 yields exact recovery under the remaining hypotheses of that theorem. This exact recovery agrees with the global profile when the unconstrained affine profile is feasible and the tangent parameterization has full parameter rank. The hierarchy also predicts the failure mode that we observe in the non-affine benchmark: Far from the nominal model, the Taylor remainder $H_\mu\|\Delta\|^2$ need not be small and the local smoothness constants need not remain uniform. The local method can then lose the second-order advantage while global profiling retains that advantage.

D.10 Affine Profiles and Training Algorithm

For residuals affine in the nuisance parameter,

$$R_\mu(u) \coloneqq R_0(u) + J(u)\mu, \qquad J(u) = \partial_\mu R_\mu(u),$$

the inner profile is a small least-squares problem. For a scalar interval $[\mu_{\min}, \mu_{\max}]$ and $J(u) \neq 0$,

$$\mu_{\mathrm{prof}}(u) = \mathrm{clip}_{[\mu_{\min},\mu_{\max}]}\left(-\frac{\langle R_0(u), J(u)\rangle}{\|J(u)\|^2}\right).$$

The implementation replaces the scalar denominator by $\max\{\|J(u)\|^2, 10^{-12}\}$. For vector parameters, the implementation enumerates active lower and upper constraints and solves each feasible free least-squares subproblem with a $10^{-8}I$ diagonal stabilization. Thus the closed form above is the exact mathematical profile, while the implementation uses a numerically stabilized counterpart. In these nuisance-affine experiments, the ideal global and tangent profile losses coincide when the unconstrained least-squares profile lies in the uncertainty set and the tangent parameterization spans the full parameter space, as holds whenever Σ is nonsingular; the implemented losses agree up to the stated numerical safeguards.

At a smooth exact interior profile, the chain rule and the inner first-order condition give the envelope derivative: The term containing the derivative of the minimizing parameter vanishes. Thus differentiating through the profile and holding the selected parameter fixed give the same first derivative of the profiled value. This gradient equality holds for an exact numerical solve as well as a closed form; implicit differentiation is needed when the profile sensitivity itself is required. For an exact constrained profile over a fixed set, the same value-gradient conclusion holds wherever the value is differentiable.

The field-valued Darcy implementation uses a ridge solve followed by radial clipping and a detached predictor update. When clipping is inactive and the ridge system is solved exactly, this update is the envelope gradient of a regularized compatibility loss. Active radial clipping need not minimize that loss over the coefficient ball. App. D.11 separates the regularization effect from inner optimization error and gives a conditional gradient-error bound. Proposition 5 instead concerns predictor error under uniform inner value accuracy and outer suboptimality in the jointly affine setting; neither statement is a convergence guarantee for detached neural-network training.

Listing 1 gives the scalar PyTorch implementation used by the nuisance-affine compatibility losses. The code writes the residual as $R_\mu = a - \mu b$; in the notation above, $a = R_0(u)$ and $b = -J(u)$. The inner profile is selected from detached tensors, after which the selected residual is formed from the live tensors a and b. Thus automatic differentiation computes the value-function derivative described above rather than optimizing μ_{prof} as a model parameter. The returned profile is retained only as a diagnostic. Except when the denominator safeguard is active, the clipped quotient is the exact minimizer over the scalar interval.

Algorithm 1 Compatibility Training

1: Choose model f_θ, nominal parameter $\hat{\mu}$, uncertainty set $\mathcal{U}$, and weight λ.
2: **for** training iteration $t = 1, \ldots, T$ **do**
3: Sample data batch and collocation batch.
4: Compute $\mathcal{L}_{\text{data}}(f_\theta)$.
5: Solve $\mu_{\text{prof}} \in \arg\min_{\mu \in \mathcal{U}} \|R_\mu(f_\theta)\|^2$ on the collocation batch.
6: Update θ using $\mathcal{L}_{\text{data}}(f_\theta) + \lambda \|R_{\mu_{\text{prof}}}(f_\theta)\|^2$.
7: **end for**

Listing 1: Scalar affine compatibility loss in PyTorch.

```
import torch

def scalar_profile_loss(a, b, interval, eps=1e-12):
    # Profile mean((a - mu * b)**2) over an interval.
    lo, hi = interval
    with torch.no_grad():
        a_det, b_det = a.detach(), b.detach()
        numerator = (a_det * b_det).mean()
        denominator = b_det.square().mean().clamp_min(eps)
        mu_prof = (numerator / denominator).clamp(lo, hi)

    residual = a - mu_prof * b
    return residual.square().mean(), mu_prof.detach()

physics_loss, mu_prof = scalar_profile_loss(a, b, bounds)
loss = data_loss + physics_weight * physics_loss
```

Algorithm 1 presents the exact-profile formulation; the finite-grid and regularized implementations are described in App. D.11.

D.11 Regularized and Inexact Profiles

The implemented profiles use finite candidate sets or stabilized coefficient solves. The following results connect these computations to the ideal profile: Grid spacing controls the approximation of a continuous family, while ridge shrinkage and inner optimality residuals control distinct errors in field profiles. Proofs appear in App. E.14.

Finite candidate sets. For a scalar interval $I = [a, b]$ with $a < b$, let G_K contain $K \geq 2$ equally spaced points including both endpoints, and set $h = (b-a)/(K-1)$. For a fixed predictor u, write $f(u, \beta) = \|R_\beta(u)\|^2$, $\mathcal{C}_I(u) = \min_{\beta \in I} f(u, \beta)$, and $\mathcal{C}_{G_K}(u) = \min_{\beta \in G_K} f(u, \beta)$.

Lemma 4 (Grid-profile value error)**.** *If $f(u, \cdot)$ is L_β-Lipschitz on I, then*

$$0 \leq \mathcal{C}_{G_K}(u) - \mathcal{C}_I(u) \leq L_\beta h/2. \tag{23}$$

If $f(u, \cdot)$ is also twice continuously differentiable near I and $|\partial^2_{\beta\beta} f(u, \beta)| \leq H_\beta$ on I, the upper bound improves to $H_\beta h^2/8$.

The second bound uses a stationary interior minimizer; if an endpoint is a minimizer, the grid error is zero. The constants may depend on u; uniform constants give uniform value bounds over a predictor set. The non-affine experiment uses $K = 25$ and $h = 2 \log \rho / 24$. The grid candidate evaluations reuse the prediction and PDE derivatives, and only the selected residual is differentiated. At a unique selected candidate, wherever the candidate losses are differentiable in u, the detached update is the exact gradient of the finite-grid objective $\mathcal{C}_{G_K}$. Lemma 4 bounds the difference between the finite-grid and continuous profile values. For a fixed interval and predictor, evaluating K candidates costs $O(Kn)$ arithmetic operations, where n is the number of residual entries, while the profile-value error bound decreases as $O((K-1)^{-1})$, or $O((K-1)^{-2})$ under the stated smoothness assumptions. The quadratic bound requires, in particular, the absence of an exponent-clipping kink on the interval.

Regularized affine profiles. Let $\mathcal{N} \subseteq \mathcal{H}$ be open, and let $r : \mathcal{N} \to \mathcal{Y}$ and $J : \mathcal{N} \to L(\mathbb{R}^p, \mathcal{Y})$ be continuously Fréchet differentiable. Define

$$f(u,\alpha) \coloneqq \|r(u) + J(u)\alpha\|^2, \qquad \Phi_\tau(u,\alpha) \coloneqq f(u,\alpha) + \tau\|\alpha\|_2^2,$$
$$\mathcal{C}_{\mathcal{U},\tau}(u) \coloneqq \min_{\alpha\in\mathcal{U}} \Phi_\tau(u,\alpha), \qquad \mathcal{U} = \overline{B}_\rho(0) \subset \mathbb{R}^p,$$

where $\rho > 0$ and $\tau > 0$ are fixed independently of u. Set $\mathcal{C}_{\mathcal{U}}(u) = \min_{\alpha\in\mathcal{U}} f(u,\alpha)$. The residual norm includes the empirical normalization: For the Darcy matrices in App. F, take $r = -y/\sqrt{n}$ and $J = X/\sqrt{n}$, so the implemented system is $J^*J + \tau I$.

Proposition 6 (Ridge profile and envelope gradient)**.** *In the setting above, the unconstrained regularized minimizer and the corresponding objective value are*

$$\alpha_\tau^{\mathrm{free}} = -(J^*J + \tau I)^{-1}J^*r, \qquad Q_\tau \coloneqq I - J(J^*J + \tau I)^{-1}J^*, \tag{24}$$
$$\mathcal{C}_{\mathbb{R}^p,\tau}(u) = \langle r, Q_\tau r\rangle, \qquad r + J\alpha_\tau^{\mathrm{free}} = Q_\tau r. \tag{25}$$

The constrained minimizer $\alpha_{\rho,\tau}(u)$ is unique, the constrained value is continuously differentiable, and

$$0 \le \mathcal{C}_{\mathcal{U},\tau}(u) - \mathcal{C}_{\mathcal{U}}(u) \le \tau\rho^2, \tag{26}$$
$$\nabla_u \mathcal{C}_{\mathcal{U},\tau}(u) = \nabla_u f(u, \alpha_{\rho,\tau}(u)), \tag{27}$$

where the derivative on the right holds the coefficient fixed. If $\alpha_\tau^{\mathrm{free}} \in \mathcal{U}$, the free minimizer equals $\alpha_{\rho,\tau}$.

Ridge gives a continuous version of nuisance removal. On a left singular direction of J with singular value $s_j > 0$, Q_τ retains the fraction $\tau/(s_j^2 + \tau)$ and is the identity on $\mathrm{Range}(J)^\perp$. As $\tau \to 0$, Q_τ approaches the unregularized projector Q at fixed J, but Q_τ is generally not idempotent for positive τ. Thus the residual alone has value $\|Q_\tau r\|^2$, whereas the regularized objective has value $\langle r, Q_\tau r\rangle$. Ridge stabilizes weak coefficient directions while retaining some represented mismatch. The ridge scale must be interpreted relative to J^*J and the basis normalization; the exact zero-penalty recovery statement for unregularized compatibility training does not transfer automatically. Both r and J may depend on u in Eq. 27, and both are differentiated with the coefficient fixed.

Accuracy of a feasible computed coefficient. For $\widehat{\alpha} \in \mathcal{U}$, define the regularized value error and inner optimality residual of the computed coefficient by

$$\eta_\tau(u) \coloneqq \Phi_\tau(u,\widehat{\alpha}) - \mathcal{C}_{\mathcal{U},\tau}(u),$$
$$\kappa_{\mathrm{in}}(u) \coloneqq \mathrm{dist}(0, \nabla_\alpha \Phi_\tau(u,\widehat{\alpha}) + N_{\mathcal{U}}(\widehat{\alpha})),$$

where $N_{\mathcal{U}}(\alpha) = \{z : \langle z, \beta - \alpha\rangle \le 0 \text{ for all } \beta \in \mathcal{U}\}$ is the outward normal cone.

Proposition 7 (Value and gradient errors of an inexact ridge profile)**.** *Under the hypotheses of Proposition 6, every feasible $\widehat{\alpha}$ satisfies*

$$\eta_\tau(u) \le \frac{\kappa_{\mathrm{in}}(u)^2}{4\tau}, \qquad \|\widehat{\alpha} - \alpha_{\rho,\tau}(u)\|_2 \le \frac{\kappa_{\mathrm{in}}(u)}{2\tau}, \tag{28}$$
$$0 \le f(u,\widehat{\alpha}) - \mathcal{C}_{\mathcal{U}}(u) \le \tau\rho^2 + \eta_\tau(u). \tag{29}$$

If, at this u, $\nabla_u f(u,\cdot)$ is $L_{u\alpha}$-Lipschitz on $\mathcal{U}$, then the detached update $g_{\mathrm{det}} \coloneqq \nabla_u f(u,\widehat{\alpha})$ satisfies

$$\|g_{\mathrm{det}} - \nabla_u \mathcal{C}_{\mathcal{U},\tau}(u)\| \le \frac{L_{u\alpha}}{2\tau}\kappa_{\mathrm{in}}(u). \tag{30}$$

For the coefficient ball, the normal cone is $\{0\}$ at interior points and $\{t\widehat{\alpha} : t \ge 0\}$ at boundary points, so κ_{in} is computable from the inner gradient and a scalar minimization. Including this normal term distinguishes a constrained optimum from a nonzero unconstrained gradient at the boundary. An exact constrained ridge solution has $\kappa_{\mathrm{in}} = 0$ and gives the exact envelope gradient even at an active constraint.

In Darcy, inactive radial clipping identifies the computed coefficient with $\alpha_\tau^{\mathrm{free}}$ up to numerical solve error. The detached update then has the regularized-envelope interpretation even though the reported loss omits $\tau\|\alpha\|^2$: That term has no explicit u-derivative at a fixed coefficient. Active

radial clipping supplies a feasible coefficient, but generally leaves nonzero κ_{in}. The bound Eq. 29 separates the optimization error of the clipped coefficient from ridge shrinkage. With the main field setting $\tau = 10^{-4}$ and $\rho = 1$, the ridge contribution is at most 10^{-4} in residual-MSE units; the computed profile can incur the additional η_τ. Averaging per-instance value bounds preserves this decomposition for an operator minibatch.

These bounds on profile approximation are deterministic error bounds for a given predictor and inner solve. The existing runs use a fixed ridge and do not log κ_{in}, so the bounds are not measured certificates for those runs. Error relative to a true coefficient, reported in App. G.6, assesses physical correction accuracy rather than inner optimality. Uniform value bounds can be used in Proposition 5 when the proposition's jointly affine and outer-suboptimality hypotheses also hold; Eq. 30 separately controls the approximation to the regularized gradient, without asserting convergence of neural-network training.

E Proofs of Formal Statements

This appendix provides complete proofs of all formal statements and the auxiliary lemmas used in the main text. The organization follows the theorem order summarized in the appendix table of contents.

Setting and conventions. Predictions u live in a real Hilbert space $\mathcal{H}$ and residuals take values in a real Hilbert space $\mathcal{Y}$; $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ denote the corresponding inner products and norms, and $\|\cdot\|$ also denotes the operator norm of a bounded linear map. Physics parameters μ live in $\mathbb{R}^p$, the uncertainty set $\mathcal{U} \subset \mathbb{R}^p$ is compact, and $\Sigma^{1/2}$ is the positive-semidefinite square root of a covariance $\Sigma \in \mathbb{R}^{p\times p}$. Derivatives in u are Fréchet derivatives $D_u R_\mu(u) : \mathcal{H} \to \mathcal{Y}$ with Hilbert adjoint $D_u R_\mu(u)^* : \mathcal{Y} \to \mathcal{H}$, and $\partial_\mu R_\mu(u) : \mathbb{R}^p \to \mathcal{Y}$ denotes the derivative in the parameter. Whenever $u \mapsto R_\mu(u)$ is Fréchet differentiable, the chain rule gives

$$\nabla_u \|R_\mu(u)\|^2 = 2 D_u R_\mu(u)^* R_\mu(u).$$

E.1 First-Order Size of the Nominal Physics Gradient

Lemma 1 is stated in App. C; we prove the lemma here.

Proof of Lemma 1. Define $F : [0,1] \to \mathcal{Y}$ by $F(s) := R_{\mu^\star + s\Delta_\mu}(u^\star)$. Choose $r > 0$ so that the closed ball $\overline{B}(\mu^\star, r)$ lies in N. Continuity of the second derivative and local Lipschitz continuity of $D_u R_\mu(u^\star)$ provide constants $H, C_{\text{Lip}} < \infty$ on this ball. For every sufficiently close $\hat{\mu}$, the segment from $\hat{\mu}$ to $\mu^\star$ lies in the ball. Taylor's theorem with integral remainder applied to F above gives

$$\begin{aligned} R_{\hat{\mu}}(u^\star) = F(1) &= F(0) + F'(0) + \int_0^1 (1-s) F''(s) ds \\ &= \partial_\mu R_{\mu^\star}(u^\star)\Delta_\mu + \rho, \qquad \|\rho\| \le \tfrac{1}{2} H \|\Delta_\mu\|^2, \end{aligned}$$

since $F''(s) = \partial^2_{\mu\mu} R_{\mu^\star + s\Delta_\mu}(u^\star)[\Delta_\mu, \Delta_\mu]$. For the gradient, write $D_u R_{\hat{\mu}}(u^\star)^* = D_u R_{\mu^\star}(u^\star)^* + E$ with $\|E\| \le C_{\text{Lip}} \|\Delta_\mu\|$ by the Lipschitz hypothesis, since adjoints preserve operator norms. Combining,

$$\begin{aligned} \nabla_u P_{\hat{\mu}}(u^\star) &= 2(D_u R_{\mu^\star}(u^\star)^* + E)(\partial_\mu R_{\mu^\star}(u^\star)\Delta_\mu + \rho) \\ &= 2 D_u R_{\mu^\star}(u^\star)^* \partial_\mu R_{\mu^\star}(u^\star)\Delta_\mu + O(\|\Delta_\mu\|^2), \end{aligned}$$

because each cross term is bounded by a constant multiple of $\|\Delta_\mu\|^2$. □

E.2 Nominal Residual Bias

Proof of Proposition 1. Write $u_\lambda := u_{\lambda,\hat{\mu}} \in \mathcal{N}$ and abbreviate $\mathcal{L} := \mathcal{L}_{\text{data}}$ and $P := P_{\hat{\mu}}$.

We begin with the first-order condition. The objective $\mathcal{J}_{\text{nom}} = \mathcal{L} + \lambda P$ is Fréchet differentiable on the open set $\mathcal{N}$ and u_λ is a local minimizer lying in $\mathcal{N}$ by hypothesis. For any $h \in \mathcal{H}$, the

scalar function $t \mapsto \mathcal{J}_{\mathrm{nom}}(u_\lambda + th)$ has a local minimum at $t = 0$ and is differentiable there, so the derivative $\langle \nabla \mathcal{J}_{\mathrm{nom}}(u_\lambda), h\rangle$ vanishes. As h is arbitrary,

$$\nabla \mathcal{L}(u_\lambda) = -\lambda \nabla P(u_\lambda). \tag{31}$$

The data gradient is strongly monotone. Since $\mathcal{N}$ is convex and $\mathcal{L}$ is m-strongly convex and differentiable on $\mathcal{N}$, for all $u, v \in \mathcal{N}$,

$$\mathcal{L}(v) \geq \mathcal{L}(u) + \langle \nabla \mathcal{L}(u), v - u\rangle + \tfrac{m}{2}\|v - u\|^2.$$

Adding this inequality to the same inequality with u and v exchanged gives

$$\langle \nabla \mathcal{L}(u) - \nabla \mathcal{L}(v), u - v\rangle \geq m\|u - v\|^2, \qquad u, v \in \mathcal{N}. \tag{32}$$

We can now assemble the bound. Apply Eq. 32 with $u = u_\lambda$ and $v = u^\star$ and use $\nabla \mathcal{L}(u^\star) = 0$ and Eq. 31:

$$m\|u_\lambda - u^\star\|^2 \leq \langle \nabla \mathcal{L}(u_\lambda), u_\lambda - u^\star\rangle = -\lambda \langle \nabla P(u_\lambda), u_\lambda - u^\star\rangle.$$

Splitting $\nabla P(u_\lambda) = \nabla P(u^\star) + (\nabla P(u_\lambda) - \nabla P(u^\star))$ and applying the Cauchy–Schwarz inequality and the K-Lipschitz property of ∇P on $\mathcal{N}$,

$$m\|u_\lambda - u^\star\|^2 \leq \lambda \|\nabla P(u^\star)\| \|u_\lambda - u^\star\| + \lambda K \|u_\lambda - u^\star\|^2.$$

Since $\lambda K < m$, rearranging gives $(m - \lambda K)\|u_\lambda - u^\star\|^2 \leq \lambda \|\nabla P(u^\star)\| \|u_\lambda - u^\star\|$. If $u_\lambda = u^\star$ the claimed bound holds trivially; otherwise dividing by $\|u_\lambda - u^\star\| > 0$ yields

$$\|u_\lambda - u^\star\| \leq \frac{\lambda}{m - \lambda K} \|\nabla P_{\hat{\mu}}(u^\star)\|.$$

□

E.3 Stationary-Point Separation and First-Order Nominal Bias

Proof of Theorem 1. Write $u_\lambda \coloneqq u_{\lambda,\hat{\mu}}$, $d \coloneqq \|u_\lambda - u^\star\|$, $\mathcal{L} \coloneqq \mathcal{L}_{\mathrm{data}}$, and $P \coloneqq P_{\hat{\mu}}$.

We begin with the first-order condition. The objective $\mathcal{J}_{\mathrm{nom}} = \mathcal{L} + \lambda P$ is Fréchet differentiable on the open set $\mathcal{N}$ by the data-loss and physics-penalty hypotheses. A stationary point satisfies $\nabla \mathcal{J}_{\mathrm{nom}}(u_\lambda) = 0$ by definition, and a local minimizer is stationary by the directional-derivative argument opening the proof of Proposition 1. In either case,

$$\nabla \mathcal{L}(u_\lambda) = -\lambda \nabla P(u_\lambda), \qquad \text{hence} \qquad \|\nabla \mathcal{L}(u_\lambda)\| = \lambda \|\nabla P(u_\lambda)\|. \tag{33}$$

Smoothness bounds the data gradient from above. By the M-Lipschitz property of $\nabla \mathcal{L}$ on $\mathcal{N}$ and $\nabla \mathcal{L}(u^\star) = 0$ from the data-loss hypothesis,

$$\|\nabla \mathcal{L}(u_\lambda)\| = \|\nabla \mathcal{L}(u_\lambda) - \nabla \mathcal{L}(u^\star)\| \leq Md. \tag{34}$$

Nondegeneracy bounds the physics gradient from below. By the K-Lipschitz property of ∇P on $\mathcal{N}$ and the reverse triangle inequality,

$$\|\nabla P(u_\lambda)\| \geq \|\nabla P(u^\star)\| - \|\nabla P(u_\lambda) - \nabla P(u^\star)\| \geq g - Kd. \tag{35}$$

A case analysis now completes the estimate. Suppose first that $Kd \geq g$. Since $g > 0$, that inequality forces $K > 0$ and $d \geq g/K$. Because $\lambda, M > 0$,

$$\frac{\lambda g}{M + \lambda K} \leq \frac{\lambda g}{\lambda K} = \frac{g}{K} \leq d,$$

which is Eq. 1. Suppose instead that $Kd < g$. Then $g - Kd > 0$. Combining the first-order condition in Eq. 33, the smoothness bound in Eq. 34, and the reverse-triangle bound in Eq. 35 gives

$$\lambda(g - Kd) \leq \lambda \|\nabla P(u_\lambda)\| = \|\nabla \mathcal{L}(u_\lambda)\| \leq Md,$$

which is equivalent to $\lambda g \leq (M + \lambda K)d$ and again yields Eq. 1. For the monotonicity claim, the function $\lambda \mapsto \lambda g/(M + \lambda K)$ is differentiable on $(0, \infty)$ with derivative $gM/(M + \lambda K)^2 > 0$ because $g, M > 0$, so the map is strictly increasing; when $K > 0$ the supremum is $\lim_{\lambda \to \infty} \lambda g/(M + \lambda K) = g/K$. □

Proof of Corollary 1. The chain-rule identity recorded above gives $\nabla_u P(u^\star) = 2A^* R_{\hat{\mu}}(u^\star)$ with $A = D_u R_{\hat{\mu}}(u^\star)$. Therefore, by the two lower-bound assumptions of the corollary,

$$g = \|\nabla_u P(u^\star)\| = 2\|A^* R_{\hat{\mu}}(u^\star)\| \geq 2\nu \|R_{\hat{\mu}}(u^\star)\| \geq 2\nu c_R \|\Delta_\mu\|.$$

Inserting this bound into Eq. 1 completes the proof. □

Stationarity and the bias sandwich. Only stationarity of u_λ is used, so the lower bound applies to any interior critical point reached by training, not only to local minimizers, and no convexity of the data term is needed. The data-loss hypothesis is a smoothness condition, satisfied in particular whenever $\mathcal{L}_{\mathrm{data}}$ is m-strongly convex and M-smooth with $m \le M$, in which case Proposition 1 and Theorem 1 together sandwich the bias,

$$\frac{\lambda g}{M + \lambda K} \le \|u_{\lambda,\hat{\mu}} - u^\star\| \le \frac{\lambda g}{m - \lambda K} \qquad \text{for } \lambda K < m.$$

A sufficient adjoint lower bound. The adjoint lower-bound assumption of Corollary 1 holds with $\nu = 1/\|A^{-1}\|$ whenever A is boundedly invertible: In that case, A^* is boundedly invertible with $(A^*)^{-1} = (A^{-1})^*$ and $\|(A^*)^{-1}\| = \|A^{-1}\|$, so for every $y \in \mathcal{Y}$,

$$\|y\| = \|(A^*)^{-1}A^*y\| \le \|A^{-1}\|\|A^*y\|, \qquad \text{equivalently,} \qquad \|A^*y\| \ge \frac{\|y\|}{\|A^{-1}\|}.$$

Bounded invertibility of A is well-posedness of the linearized nominal equation after the relevant boundary or initial conditions have been built into the domain and codomain of A. Such invertibility is a sufficient example for the adjoint lower bound, not an automatic consequence of the equation being elliptic or parabolic.

E.4 Bias under Uncertainty-Set Exclusion

The next lemma states and proves the distance-to-set bounds referred to in Sec. D.1.

Lemma 5 (Distance-to-set upper and gradient bounds)**.** *Let $\mathcal{U} \subset \mathbb{R}^p$ be compact and let $R_{\mu^\star}(u^\star) = 0$. If $\mu \mapsto R_\mu(u^\star)$ is L_μ-Lipschitz on a set containing $\mathcal{U} \cup \{\mu^\star\}$, then the profiled value obeys the upper bound*

$$\mathcal{C}_{\mathcal{U}}(u^\star) \le L_\mu^2 \operatorname{dist}(\mu^\star,\mathcal{U})^2. \tag{36}$$

If in addition $\sup_{\mu\in\mathcal{U}} \|D_u R_\mu(u^\star)\| \le B$ and each map $u \mapsto R_\mu(u)$ is Fréchet differentiable at $u^\star$, then every active profile parameter $\bar{\mu} \in \arg\min_{\mu\in\mathcal{U}} \|R_\mu(u^\star)\|^2$ satisfies the gradient bound

$$\|\nabla g_{\bar{\mu}}(u^\star)\| \le 2B\sqrt{\mathcal{C}_{\mathcal{U}}(u^\star)} \le 2BL_\mu \operatorname{dist}(\mu^\star,\mathcal{U}), \qquad g_{\bar{\mu}} \coloneqq \|R_{\bar{\mu}}(\cdot)\|^2. \tag{37}$$

Proof of Lemma 5. For Eq. 36, the function $\mu \mapsto \|\mu - \mu^\star\|$ is continuous on the compact set $\mathcal{U}$, so there is a nearest point $\mu_{\mathcal{U}} \in \mathcal{U}$ with $\|\mu_{\mathcal{U}} - \mu^\star\| = \operatorname{dist}(\mu^\star,\mathcal{U})$. Using $R_{\mu^\star}(u^\star) = 0$ and the Lipschitz property,

$$\mathcal{C}_{\mathcal{U}}(u^\star) \le \|R_{\mu_{\mathcal{U}}}(u^\star)\|^2 = \|R_{\mu_{\mathcal{U}}}(u^\star) - R_{\mu^\star}(u^\star)\|^2 \le L_\mu^2\|\mu_{\mathcal{U}} - \mu^\star\|^2 = L_\mu^2 \operatorname{dist}(\mu^\star,\mathcal{U})^2.$$

For Eq. 37, the chain-rule identity recorded above applied at the parameter $\bar{\mu}$ gives

$$\nabla g_{\bar{\mu}}(u^\star) = 2D_u R_{\bar{\mu}}(u^\star)^* R_{\bar{\mu}}(u^\star).$$

Consequently,

$$\|\nabla g_{\bar{\mu}}(u^\star)\| \le 2\|D_u R_{\bar{\mu}}(u^\star)\|\|R_{\bar{\mu}}(u^\star)\| \le 2B\sqrt{\mathcal{C}_{\mathcal{U}}(u^\star)},$$

because $\bar{\mu}$ attains the minimum, so that $\|R_{\bar{\mu}}(u^\star)\|^2 = \mathcal{C}_{\mathcal{U}}(u^\star)$. Combining the active-branch gradient bound with Eq. 36 gives the second inequality. If $\mathcal{C}_{\mathcal{U}}$ is differentiable at $u^\star$, the same bound applies to the envelope gradient without requiring a unique active parameter. Indeed, for every active $\bar{\mu}$, the smooth branch $g_{\bar{\mu}}$ of Eq. 37 majorizes $\mathcal{C}_{\mathcal{U}}(u)$ and touches the envelope at $u^\star$. Hence the differentiable function $g_{\bar{\mu}} - \mathcal{C}_{\mathcal{U}}$ has a local minimum at $u^\star$, so $\nabla\mathcal{C}_{\mathcal{U}}(u^\star) = \nabla g_{\bar{\mu}}(u^\star)$. □

E.5 Equivalence to Joint Inverse Optimization

Proof of Proposition 2. Write $G(\theta) \coloneqq \mathcal{J}_{\mathrm{comp}}(f_\theta)$ and let $M(\theta)$ be as in Sec. D.2. Since the data term is independent of μ and $\lambda > 0$, $G(\theta) = \min_{\mu\in\mathcal{U}} F(\theta,\mu)$. For every pair (θ,μ), $F(\theta,\mu) \ge G(\theta) \ge \inf_\vartheta G(\vartheta)$, so

$$\inf_{\theta,\mu\in\mathcal{U}} F(\theta,\mu) \ge \inf_\theta G(\theta).$$

Conversely, choose any $\mu_\theta \in M(\theta)$ for a fixed θ. Then $\inf_{\vartheta,\nu} F(\vartheta,\nu) \le F(\theta,\mu_\theta) = G(\theta)$; taking the infimum over θ gives the reverse inequality.

If θ minimizes G and $\mu \in M(\theta)$, the pair attains this common value and minimizes F. If (θ,μ) minimizes F, then μ must minimize the inner problem at θ, since otherwise replacing μ would strictly lower F. The infimum identity then implies that θ minimizes G. This minimizer correspondence proves Eq. 7 and, by applying $\theta \mapsto f_\theta$, equality of the optimal predictor sets. The same infimum identity without attainment follows directly from $\inf_\theta \inf_\mu F = \inf_{\theta,\mu} F$.

For Eq. 8, each summand is bounded below by the minimum over the corresponding nuisance set. Choosing a minimizer independently in each of the finitely many sets $\mathcal{U}_i$ attains the sum of these lower bounds, proving equality. When all $\mathcal{U}_i = \mathcal{U}$, imposing $\mu_i = \mu$ restricts the product set to the diagonal where all instance parameters agree; that shared-parameter minimum can therefore be larger. □

E.6 Robust Recovery by Compatibility Training

Under the hypotheses of Theorem 2, we abbreviate $\mathcal{L} \coloneqq \mathcal{L}_{\mathrm{data}}$, and write

$$c(u) \coloneqq \min_{\mu\in\mathcal{U}} \|R_\mu(u)\|, \qquad \mathcal{C}_{\mathcal{U}}(u) = c(u)^2,$$

where the minimum is attained for every $u \in \mathcal{N}$ because $\mu \mapsto \|R_\mu(u)\|^2$ is continuous on the compact set $\mathcal{U}$ by the residual-continuity assumption. Since $\mathcal{N}$ is convex, the mean value inequality and the uniform residual-derivative bound give the uniform Lipschitz property

$$\|R_\mu(u) - R_\mu(v)\| \le B\|u - v\|, \qquad u, v \in \mathcal{N}, \quad \mu \in \mathcal{U}. \tag{38}$$

We also record that, by the chain-rule identity above, for every $\mu \in \mathcal{U}$, the function $u \mapsto \|R_\mu(u)\|^2$ is Fréchet differentiable on $\mathcal{N}$ with gradient $2D_u R_\mu(u)^* R_\mu(u)$.

Let $\mu_v \in \mathcal{U}$ attain $c(v)$. Then, by Eq. 38,

$$c(u) \le \|R_{\mu_v}(u)\| \le \|R_{\mu_v}(v)\| + B\|u - v\| = c(v) + B\|u - v\|.$$

Exchanging the roles of u and v gives

$$|c(u) - c(v)| \le B\|u - v\|, \qquad u, v \in \mathcal{N}. \tag{39}$$

Lemma 6 (Exact first-order condition at compatibility minimizers)**.** *Let $u_\lambda \in \mathcal{N}$ be a local minimizer of $\mathcal{J}_{\mathrm{comp}}$ and let $\mu_\lambda \in \arg\min_{\mu\in\mathcal{U}} \|R_\mu(u_\lambda)\|^2$. Then*

$$\nabla\mathcal{L}(u_\lambda) = -2\lambda D_u R_{\mu_\lambda}(u_\lambda)^* R_{\mu_\lambda}(u_\lambda), \qquad \|\nabla\mathcal{L}(u_\lambda)\| \le 2\lambda B c(u_\lambda).$$

Proof of Lemma 6. Define the auxiliary function $\varphi(u) \coloneqq \mathcal{L}(u) + \lambda\|R_{\mu_\lambda}(u)\|^2$ on $\mathcal{N}$, with the profiled parameter frozen at μ_λ. For every $u \in \mathcal{N}$,

$$\|R_{\mu_\lambda}(u)\|^2 \ge \min_{\mu\in\mathcal{U}} \|R_\mu(u)\|^2 = \mathcal{C}_{\mathcal{U}}(u),$$

so $\varphi \ge \mathcal{J}_{\mathrm{comp}}$ pointwise on $\mathcal{N}$, while $\varphi(u_\lambda) = \mathcal{J}_{\mathrm{comp}}(u_\lambda)$ because μ_λ attains the profile at u_λ. Let $r_0 > 0$ be such that $\{u : \|u - u_\lambda\| < r_0\} \subseteq \mathcal{N}$ and $\mathcal{J}_{\mathrm{comp}}(u) \ge \mathcal{J}_{\mathrm{comp}}(u_\lambda)$ for all u with $\|u - u_\lambda\| < r_0$. For those u,

$$\varphi(u) \ge \mathcal{J}_{\mathrm{comp}}(u) \ge \mathcal{J}_{\mathrm{comp}}(u_\lambda) = \varphi(u_\lambda),$$

so the smooth majorant φ, which touches $\mathcal{J}_{\mathrm{comp}}$ at u_λ, also has a local minimum at u_λ. By the data-loss assumption and the differentiability of $u \mapsto \|R_{\mu_\lambda}(u)\|^2$ recorded above, φ is Fréchet differentiable on $\mathcal{N}$ with $\nabla\varphi(u) = \nabla\mathcal{L}(u) + 2\lambda D_u R_{\mu_\lambda}(u)^* R_{\mu_\lambda}(u)$. For every $h \in \mathcal{H}$, the function $t \mapsto \varphi(u_\lambda + th)$ is differentiable at $t = 0$ and has a local minimum there, so $\langle\nabla\varphi(u_\lambda), h\rangle = 0$; as h is arbitrary, $\nabla\varphi(u_\lambda) = 0$, which is the stated identity. The bound follows from $\|A^*\| = \|A\|$ for bounded linear A, the uniform residual-derivative bound, and $\|R_{\mu_\lambda}(u_\lambda)\| = c(u_\lambda)$. □

Lemma 6 does not require $\mathcal{C}_{\mathcal{U}}$ to be differentiable at u_λ, and does not require the profile minimizer μ_λ to be unique: Any attained profile parameter yields the same conclusion. The smooth-majorant argument replaces the envelope differentiation of Danskin type applied to $\mathcal{C}_{\mathcal{U}}$ by an elementary comparison that is valid in the infinite-dimensional setting used here.

Proof of Theorem 2. Let $u_\lambda \in \mathcal{N}$ be a local minimizer of $\mathcal{J}_{\mathrm{comp}}$ and write $d := \|u_\lambda - u^\star\|$. If $d = 0$ the bound holds trivially, so assume $d > 0$.

We first bound the data gradient at the minimizer. By Lemma 6 and Eq. 39,

$$\|\nabla\mathcal{L}(u_\lambda)\| \le 2\lambda B c(u_\lambda) \le 2\lambda B(c(u^\star) + Bd).$$

We next bound the profiled residual at the truth. The true-model and parameter-Lipschitz assumptions are exactly the assumptions used to derive Eq. 36 in Lemma 5, so $\mathcal{C}_{\mathcal{U}}(u^\star) \le L_\mu^2 \operatorname{dist}(\mu^\star, \mathcal{U})^2$, and hence $c(u^\star) \le L_\mu \operatorname{dist}(\mu^\star, \mathcal{U})$.

Strong monotonicity then closes the argument. Adding the strong-convexity inequality on $\mathcal{N}$ at the ordered pair $(u_\lambda, u^\star)$ to the same inequality with the roles exchanged gives, exactly as in Eq. 32,

$$md^2 \le \langle \nabla\mathcal{L}(u_\lambda) - \nabla\mathcal{L}(u^\star), u_\lambda - u^\star\rangle \le (\|\nabla\mathcal{L}(u_\lambda)\| + \varepsilon)d,$$

using the Cauchy–Schwarz inequality and $\|\nabla\mathcal{L}(u^\star)\| \le \varepsilon$ from the data-loss assumption. Dividing by $d > 0$ and inserting the two preceding bounds,

$$md \le \varepsilon + 2\lambda B L_\mu \operatorname{dist}(\mu^\star, \mathcal{U}) + 2\lambda B^2 d.$$

Since $2B^2\lambda < m$, rearranging gives Eq. 4. □

Proof of Corollary 2. Write $c(u) := \min_{\mu\in\mathcal{U}} \|R_\mu(u)\|$ and $d := \|u_\lambda - u^\star\|$. Compactness and continuity ensure attainment. Convexity of $\mathcal{N}$ and the uniform derivative bound give $|c(u) - c(v)| \le B\|u - v\|$, by the candidate-profile comparison used in Eq. 39. For any active parameter μ_λ at the local minimizer, the differentiable function $\mathcal{L}_{\mathrm{data}}(u) + \lambda\|R_{\mu_\lambda}(u)\|^2$ majorizes $\mathcal{J}_{\mathrm{comp}}$ and agrees with $\mathcal{J}_{\mathrm{comp}}$ at u_λ. The smooth majorant therefore has a local minimum at u_λ, exactly as in Lemma 6. Hence

$$\|\nabla\mathcal{L}_{\mathrm{data}}(u_\lambda)\| \le 2\lambda B c(u_\lambda) \le 2\lambda B(\eta_{\mathcal{U}} + Bd).$$

If $d = 0$, the claim is immediate. Otherwise, strong monotonicity of the data gradient and the assumed data-gradient bound at $u^\star$ give

$$md \le \|\nabla\mathcal{L}_{\mathrm{data}}(u_\lambda)\| + \varepsilon \le 2\lambda B\eta_{\mathcal{U}} + 2\lambda B^2 d + \varepsilon.$$

Rearranging under $2\lambda B^2 < m$ proves Eq. 5. No zero-residual parameter or parameter-Lipschitz assumption is used. □

Proof of the value-comparison bound. For the closed ball $\bar{\mathcal{V}}$ and the corresponding minimizer u_λ defined in App. D.1, from $\mathcal{J}_{\mathrm{comp}}(u_\lambda) \le \mathcal{J}_{\mathrm{comp}}(u^\star)$ and $\mathcal{C}_{\mathcal{U}}(u_\lambda) \ge 0$,

$$\mathcal{L}(u_\lambda) - \mathcal{L}(u^\star) \le \lambda\mathcal{C}_{\mathcal{U}}(u^\star) \le \lambda L_\mu^2 \operatorname{dist}(\mu^\star, \mathcal{U})^2$$

by Eq. 36, while strong convexity and $\|\nabla\mathcal{L}(u^\star)\| \le \varepsilon$ give, with $d := \|u_\lambda - u^\star\|$,

$$\mathcal{L}(u_\lambda) - \mathcal{L}(u^\star) \ge \langle \nabla\mathcal{L}(u^\star), u_\lambda - u^\star\rangle + \tfrac{m}{2}d^2 \ge -\varepsilon d + \tfrac{m}{2}d^2.$$

Combining the two preceding bounds and solving the quadratic inequality $(m/2)d^2 - \varepsilon d - \lambda L_\mu^2 \operatorname{dist}(\mu^\star, \mathcal{U})^2 \le 0$ for $d \ge 0$,

$$d \le \frac{\varepsilon + \sqrt{\varepsilon^2 + 2m\lambda L_\mu^2 \operatorname{dist}(\mu^\star, \mathcal{U})^2}}{m} \le \frac{2\varepsilon}{m} + L_\mu\sqrt{\frac{2\lambda}{m}} \operatorname{dist}(\mu^\star, \mathcal{U}).$$

The second inequality is Eq. 6.

E.7 Strict Separation of Nominal and Compatibility Training

Proof of Corollary 3. Consider first the compatibility objective. Containment gives $\operatorname{dist}(\mu^\star, \mathcal{U}) = 0$, and the data-gradient assumption makes Theorem 2 applicable with $\varepsilon = 0$. Therefore every local minimizer of $\mathcal{J}_{\mathrm{comp}}$ in $\mathcal{N}$ equals $u^\star$ when $2B^2\lambda < m$. Conversely, $\mathcal{C}_{\mathcal{U}}(u^\star) = 0$ because the profile can select $\mu^\star$, while strong convexity and $\nabla\mathcal{L}_{\mathrm{data}}(u^\star) = 0$ give

$$\mathcal{J}_{\mathrm{comp}}(u) \ge \mathcal{J}_{\mathrm{comp}}(u^\star) + \tfrac{m}{2}\|u - u^\star\|^2 \qquad (u \in \mathcal{N}).$$

Thus $u^\star$ is itself a local minimizer, and the set of local minimizers is exactly $\{u^\star\}$.

Consider next the nominal objective. Every local minimizer $u \in \mathcal{N}$ of $\mathcal{J}_{\mathrm{nom}}$ satisfies the hypotheses of Theorem 1 by assumption, so Eq. 1 gives $\|u - u^\star\| \ge \lambda g/(M + \lambda K)$, and this lower bound is strictly positive because $\lambda > 0$, $g > 0$, and $M + \lambda K > 0$. □

E.8 Profile Stability

Throughout this subsection, the hypotheses of Theorem 2 and the identifiability lower bound of Proposition 3 are in force, and we freely use the notation $c(u) = \min_{\mu \in \mathcal{U}} \|R_\mu(u)\|$, the uniform Lipschitz bound in Eq. 38, and the profiled-norm bound in Eq. 39 from App. E.6.

Proof of Proposition 3. Let $u_\lambda \in \mathcal{N}$ be a local minimizer of $\mathcal{J}_{\text{comp}}$ and write $d \coloneqq \|u_\lambda - u^\star\|$. The profile at u_λ is attained because $\mu \mapsto \|R_\mu(u_\lambda)\|^2$ is continuous on the compact set $\mathcal{U}$ by the residual-continuity assumption of Theorem 2; let μ_λ be any minimizer, so that $\|R_{\mu_\lambda}(u_\lambda)\| = c(u_\lambda)$.

We first use identifiability at the truth. Since $\mu_\lambda \in \mathcal{U}$, the identifiability lower bound applies at $\mu = \mu_\lambda$ and gives

$$c_{\text{id}} \|\mu_\lambda - \mu^\star\| \le \|R_{\mu_\lambda}(u^\star)\|.$$

The residual is then transported from the truth to the trained predictor. By the uniform Lipschitz bound in Eq. 38,

$$\|R_{\mu_\lambda}(u^\star)\| \le \|R_{\mu_\lambda}(u_\lambda)\| + Bd = c(u_\lambda) + Bd.$$

Finally we bound the profiled residual at the trained predictor. By Eq. 39 and Eq. 36, whose assumptions are the true-model and parameter-Lipschitz conditions of Theorem 2,

$$c(u_\lambda) \le c(u^\star) + Bd \le L_\mu \operatorname{dist}(\mu^\star, \mathcal{U}) + Bd.$$

Combining the three preceding bounds,

$$\|\mu_\lambda - \mu^\star\| \le \frac{1}{c_{\text{id}}} (L_\mu \operatorname{dist}(\mu^\star, \mathcal{U}) + 2Bd),$$

which is Eq. 9. Bounding d with Eq. 4 in Theorem 2, which applies because $2B^2\lambda < m$, gives Eq. 10. □

E.9 Finite-Collocation Generalization

Proof of Theorem 3. For a measurable function f of a collocation point, write $Pf \coloneqq \mathbb{E} f(Z)$ and $P_{n_c} f \coloneqq n_c^{-1} \sum_{i=1}^{n_c} f(Z_i)$.

We first establish uniform concentration before profiling. Fix $\tau > 0$ and let $\mathcal{A}_\tau$ be a τ-net of $\mathcal{K}$ with cardinality $N_\tau \coloneqq \mathfrak{N}(\mathcal{K}, d_{\mathcal{K}}, \tau)$, which is finite because $\mathcal{K}$ is compact. For a fixed $a = (u_a, \mu_a) \in \mathcal{A}_\tau$, the random variables $\ell_a(Z_i)$ are independent and lie in $[0, M_c]$. Hoeffding's inequality therefore gives

$$\mathbb{P}(|(P_{n_c} - P)\ell_a| > t) \le 2 \exp\left(-\frac{2n_c t^2}{M_c^2}\right).$$

A union bound over $\mathcal{A}_\tau$ and the choice

$$t \coloneqq M_c \sqrt{\frac{\log(2N_\tau/\delta)}{2n_c}}$$

show that, with probability at least $1 - \delta$,

$$\max_{a \in \mathcal{A}_\tau} |(P_{n_c} - P)\ell_a| \le t. \tag{40}$$

Work on this event and fix $x = (u, \mu) \in \mathcal{K}$. Choose $a \in \mathcal{A}_\tau$ with $d_{\mathcal{K}}(x, a) \le \tau$. The uniform Lipschitz assumption implies both

$$|P_{n_c}\ell_x - P_{n_c}\ell_a| \le L_c\tau, \qquad |P\ell_x - P\ell_a| \le L_c\tau.$$

Combining these inequalities with Eq. 40 yields

$$|(P_{n_c} - P)\ell_x| \le 2L_c\tau + t = \eta_{n_c}(\tau, \delta).$$

Since x was arbitrary,

$$\sup_{(u,\mu)\in\mathcal{K}} |(P_{n_c} - P)\ell_{u,\mu}| \le \eta_{n_c}(\tau, \delta). \tag{41}$$

Profiling then preserves the uniform deviation. For any two real functions a, b on the same nonempty set,

$$|\inf_x a(x) - \inf_x b(x)| \le \sup_x |a(x) - b(x)|.$$

Indeed, $a(x) \le b(x) + \sup|a - b|$ for every x, so taking infima gives the upper one-sided inequality, and exchanging a and b gives the other. Apply this fact for each fixed u to $a(\mu) \coloneqq P_{n_c}\ell_{u,\mu}$ and $b(\mu) \coloneqq P\ell_{u,\mu}$. The minima exist because the Lipschitz assumption makes both functions continuous in μ and $\mathcal{U}$ is compact. Thus Eq. 41 gives

$$|\widehat{\mathcal{C}}_{n_c}(u) - \mathcal{C}(u)| \le \eta_{n_c}(\tau, \delta) \qquad (u \in \mathcal{V}),$$

which proves Eq. 12. □

E.10 Profiled Curvature and Physics-Assisted Identifiability

Proof of Lemma 2. Define

$$\Phi(u, \mu) \coloneqq \tfrac{1}{2}\|R_\mu(u)\|^2, \qquad G(u, \mu) \coloneqq \partial_\mu \Phi(u, \mu) = \partial_\mu R_\mu(u)^* R_\mu(u) \in \mathbb{R}^p.$$

The assumed regularity of R implies that G is twice continuously Fréchet differentiable near $(u^\star, \mu^\star)$. At the true pair, $R_{\mu^\star}(u^\star) = 0$, hence $G(u^\star, \mu^\star) = 0$. For $v \in \mathbb{R}^p$, the product and chain rules give

$$D_\mu G(u^\star, \mu^\star)v = D_\mu[\partial_\mu R_\mu(u)^*]_{(u^\star,\mu^\star)}[v]R_{\mu^\star}(u^\star) + J^*Jv = J^*Jv.$$

The first term vanishes because the residual at the true pair is zero. The lower bound $J^*J \succeq \sigma^2 I_p$ makes this derivative invertible. The implicit-function theorem therefore provides neighborhoods $\mathcal{N}_1$ of $u^\star$ and W of $\mu^\star$, and a unique twice continuously differentiable map $\mu_{\mathrm{loc}} : \mathcal{N}_1 \to W$ such that

$$G(u, \mu_{\mathrm{loc}}(u)) = 0, \qquad \mu_{\mathrm{loc}}(u^\star) = \mu^\star,$$

and no other zeros of G lie in $\mathcal{N}_1 \times W$.

After shrinking $\mathcal{N}_1$ and taking W to be a convex ball whose closure lies in $\operatorname{int}\mathcal{U}$, continuity of $D^2_{\mu\mu}\Phi$ and $D^2_{\mu\mu}\Phi(u^\star, \mu^\star) = J^*J$ give

$$D^2_{\mu\mu}\Phi(u, \mu)[v, v] \ge \tfrac{1}{2}\sigma^2\|v\|^2 \quad (u \in \mathcal{N}_1, \quad \mu \in W, \quad v \in \mathbb{R}^p).$$

Thus $\Phi(u, \cdot)$ is strongly convex on W. The stationary point $\mu_{\mathrm{loc}}(u)$ is consequently the unique minimizer of $\Phi(u, \cdot)$ on W.

The remaining step derives the profile derivative. Differentiate $G(u, \mu_{\mathrm{loc}}(u)) = 0$ at $u^\star$. For $h \in \mathcal{H}$, the term differentiating $\partial_\mu R_\mu(u)^*$ is again multiplied by $R_{\mu^\star}(u^\star) = 0$, so

$$D_u G(u^\star, \mu^\star)h = J^*Ah.$$

Together with $D_\mu G(u^\star, \mu^\star) = J^*J$, the chain rule yields

$$D\mu_{\mathrm{loc}}(u^\star)h = -(J^*J)^{-1}J^*Ah.$$

□

Proof of Corollary 4. We keep the notation $\Phi(u, \mu) = \|R_\mu(u)\|^2/2$ and $G(u, \mu) = \partial_\mu\Phi(u, \mu)$ of the preceding proof, and let $\mathcal{N}_1$, W, and μ_{loc} be supplied by Lemma 2. We now upgrade this local profile to the global profile over $\mathcal{U}$. Choose concentric open balls $W_0 \Subset W_1 \Subset W$ about $\mu^\star$ and shrink $\mathcal{N}_1$ so that $\mu_{\mathrm{loc}}(u) \in W_0$ for every $u \in \mathcal{N}_1$. The set $K \coloneqq \mathcal{U} \setminus W_1$ is compact. If K is empty, set $\mathcal{N}_0 \coloneqq \mathcal{N}_1$; then every global minimizer over $\mathcal{U}$ already lies in W_1. If K is nonempty, the assumed uniqueness of the compatible parameter at $u^\star$ implies

$$c \coloneqq \min_{\mu\in K} \Phi(u^\star, \mu) > 0.$$

In that case, joint continuity, compactness of K, and $\Phi(u, \mu_{\mathrm{loc}}(u)) \to \Phi(u^\star, \mu^\star) = 0$ imply, after another shrinking to a neighborhood $\mathcal{N}_0 \subseteq \mathcal{N}_1$, that

$$\inf_{\mu \in K} \Phi(u, \mu) > \tfrac{1}{2}c \quad \text{and} \quad \Phi(u, \mu_{\mathrm{loc}}(u)) < \tfrac{1}{2}c \qquad (u \in \mathcal{N}_0).$$

Thus every global minimizer over the compact set $\mathcal{U}$ lies in $W_1 \subset W$.

The remaining step identifies that minimizer with the local profile. Such a minimizer is interior to $\mathcal{U}$, hence is a zero of $G(u, \cdot)$; uniqueness of the zero in W forces that minimizer to equal $\mu_{\mathrm{loc}}(u)$. In either case, $\mu(u) = \mu_{\mathrm{loc}}(u)$ is the unique global minimizer. Hence, on $\mathcal{N}_0$,

$$\mathcal{C}_{\mathcal{U}}(u) = 2\Phi(u, \mu(u)) = \|R_{\mu(u)}(u)\|^2.$$

Since R and $\mu(\cdot)$ are twice continuously differentiable, so is $\mathcal{C}_{\mathcal{U}}$. □

Proof of Theorem 4. Let $r(u) \coloneqq R_{\mu(u)}(u)$. Then $r(u^\star) = 0$ and

$$Dr(u^\star) = A + JD\mu(u^\star) = (I - J(J^*J)^{-1}J^*)A = QA.$$

The operator $P \coloneqq J(J^*J)^{-1}J^*$ is self-adjoint and idempotent, has range $\mathrm{Range}(J)$, and acts as the identity on that range. The operator P is therefore the orthogonal projector onto $\mathrm{Range}(J)$, so $Q = I - P$ is the orthogonal projector onto the orthogonal complement of that range.

Since $\mathcal{C}_{\mathcal{U}}(u) = \langle r(u), r(u)\rangle$, the first and second derivative chain rules give

$$D\mathcal{C}_{\mathcal{U}}(u^\star)h = 2\langle r(u^\star), Dr(u^\star)h\rangle = 0$$

and, for $h, k \in \mathcal{H}$,

$$D^2\mathcal{C}_{\mathcal{U}}(u^\star)[h,k] = 2\langle Dr(u^\star)h, Dr(u^\star)k\rangle + 2\langle r(u^\star), D^2r(u^\star)[h,k]\rangle = 2\langle QAh, QAk\rangle.$$

□

Proof of Corollary 5. At $u^\star$,

$$D^2\mathcal{J}_{\mathrm{comp}}(u^\star)[h,h] = D^2\mathcal{L}_{\mathrm{data}}(u^\star)[h,h] + 2\lambda\|QAh\|^2 \geq \gamma\|h\|^2$$

by Eq. 14. The Hessian of $\mathcal{J}_{\mathrm{comp}}$ is continuous in operator norm near $u^\star$. Hence there is a convex ball $\mathcal{B} \subseteq \mathcal{N}_0$ centered at $u^\star$ on which

$$D^2\mathcal{J}_{\mathrm{comp}}(u)[h,h] \geq \tfrac{1}{2}\gamma\|h\|^2 \qquad (u \in \mathcal{B}, \quad h \in \mathcal{H}).$$

Taylor's formula with integral remainder along each segment in $\mathcal{B}$ shows that $\mathcal{J}_{\mathrm{comp}}$ is $(\gamma/2)$-strongly convex there. Moreover $\nabla\mathcal{J}_{\mathrm{comp}}(u^\star) = 0$, since both the data gradient and the compatibility gradient vanish. Strong convexity therefore gives, for every $u \in \mathcal{B}$,

$$\mathcal{J}_{\mathrm{comp}}(u) \geq \mathcal{J}_{\mathrm{comp}}(u^\star) + \tfrac{1}{4}\gamma\|u - u^\star\|^2,$$

so $u^\star$ is the unique minimizer on $\mathcal{B}$. □

Proof of Corollary 6. In finite-dimensional model coordinates with $D^2\mathcal{L}_{\mathrm{data}}(u^\star) = O^*O$, the left-hand side of Eq. 14 is the continuous nonnegative quadratic form

$$q(h) \coloneqq \|Oh\|^2 + 2\lambda\|QAh\|^2.$$

If a nonzero h lies in $\ker O \cap \ker(QA)$, no positive lower bound is possible. Conversely, if that intersection is $\{0\}$, then q is strictly positive on the compact unit sphere, so q has a positive minimum γ. This lower bound is exactly Eq. 14. □

Proof of Proposition 4. The smooth interior profile of Corollary 4 has a vanishing nuisance gradient. The chain rule therefore gives

$$\nabla\mathcal{J}_{\mathrm{comp}}(u) = \nabla_u\mathcal{J}_{\mathrm{joint}}(u, \mu(u)).$$

Thus the reduced gradient vanishes exactly when both joint gradients vanish on this minimizing branch.

At the compatible pair, differentiating the squared residual twice leaves only products of the residual's first derivatives because $R_{\mu^\star}(u^\star) = 0$. Adding the data Hessian gives Eq. 15. Set

$$K \coloneqq H_d + 2\lambda A^* A, \qquad B \coloneqq 2\lambda A^* J, \qquad D \coloneqq 2\lambda J^* J.$$

The regular-profile assumption and $\lambda > 0$ imply $D \succ 0$. The Schur complement of the nuisance block D in H_{joint} is

$$S \coloneqq K - BD^{-1}B^* = H_d + 2\lambda A^*[I - J(J^*J)^{-1}J^*]A = H_{\text{red}},$$

where the last equality is Theorem 4 with the data Hessian added. Completing the square yields, for every h, v,

$$\left\langle H_{\text{joint}} \begin{pmatrix} h \\ v \end{pmatrix}, \begin{pmatrix} h \\ v \end{pmatrix} \right\rangle = \langle Sh, h\rangle + \langle D(v + D^{-1}B^*h), v + D^{-1}B^*h\rangle.$$

The second term is minimized at $v = -D^{-1}B^*h$, proving Eq. 17. This decomposition also proves $H_{\text{joint}} \succ 0 \iff S \succ 0$.

Now suppose $H_{\text{joint}} \succ 0$, and denote the smallest and largest eigenvalues of H_{joint} by $a > 0$ and b. The quadratic form of H_{joint} is bounded below by $a(\|h\|^2 + \|v\|^2)$. Minimizing over v therefore gives $\langle Sh, h\rangle \geq a\|h\|^2$. Testing the same minimum with $v = 0$ gives $\langle Sh, h\rangle \leq b\|h\|^2$. Consequently $\lambda_{\min}(S) \geq a$ and $\lambda_{\max}(S) \leq b$, so

$$\kappa_2(S) = \frac{\lambda_{\max}(S)}{\lambda_{\min}(S)} \leq \frac{b}{a} = \kappa_2(H_{\text{joint}}).$$

All spectral comparisons use the fixed product Euclidean coordinates specified in the proposition. □

Proof of Corollary 7. Write $\Pi_i \coloneqq \Pi_{\mathcal{T}_i}$. Since $\mathcal{T}_1 \subseteq \mathcal{T}_2$ and both spaces are finite dimensional, hence closed,

$$\mathcal{T}_2 = \mathcal{T}_1 \oplus \mathcal{E}, \qquad \Pi_2 = \Pi_1 + \Pi_{\mathcal{E}}, \qquad Q_1 = Q_2 + \Pi_{\mathcal{E}},$$

where the sums are orthogonal. In particular, $Q_2 y \perp \Pi_{\mathcal{E}} z$ for every $y, z \in \mathcal{Y}$. Hence for all $h, k \in \mathcal{H}$,

$$\langle Q_1 Ah, Q_1 Ak\rangle = \langle Q_2 Ah, Q_2 Ak\rangle + \langle \Pi_{\mathcal{E}} Ah, \Pi_{\mathcal{E}} Ak\rangle. \tag{42}$$

Theorem 4, applied to each regular profile, gives

$$D^2\mathcal{C}_i(u^\star)[h, k] = 2\langle Q_i Ah, Q_i Ak\rangle.$$

Subtracting the two formulas and using Eq. 42 proves Eq. 19. Taking $k = h$ makes the right-hand side nonnegative, which is exactly the claimed positive-semidefinite order. □

Proof of Corollary 8. Orthogonality in Eq. 42 and the assumed angle bound give

$$\|Q_2 Ah\|^2 = \|Q_1 Ah\|^2 - \|\Pi_{\mathcal{E}} Ah\|^2 \geq (1 - \vartheta^2)\|Q_1 Ah\|^2.$$

Consequently,

$$q_2(h) \geq \|Oh\|^2 + 2\lambda(1 - \vartheta^2)\|Q_1 Ah\|^2 \geq (1 - \vartheta^2)q_1(h) \geq (1 - \vartheta^2)\gamma_1\|h\|^2,$$

which completes the proof. □

E.11 Global Jointly Affine Recovery

Proof of Theorem 5. Injectivity of J implies

$$\langle J^*Jv, v\rangle = \|Jv\|^2 > 0 \qquad (v \neq 0),$$

so J^*J is a positive-definite $p \times p$ matrix and is invertible. As in the projector calculation in the preceding proof, $P \coloneqq J(J^*J)^{-1}J^*$ is the orthogonal projector onto $\text{Range}(J)$ and $Q = I - P$ is the orthogonal projector onto the orthogonal complement of that range. For $z \coloneqq Au - b$, the definition of $\bar{\mu}(u)$ gives $Pz = -J\bar{\mu}(u)$ and hence, for every $\mu \in \mathbb{R}^p$,

$$z + J\mu = Qz + J(\mu - \bar{\mu}(u)).$$

The two terms on the right are orthogonal. Therefore

$$\|R_\mu(u)\|^2 = \|Q(Au-b)\|^2 + \|J(\mu - \bar{\mu}(u))\|^2. \tag{43}$$

Because $\bar{\mu}(u) \in \mathcal{U}$ for every $u \in \mathcal{V}$, the second term of Eq. 43 is minimized at zero; injectivity of J makes $\bar{\mu}(u)$ the unique minimizer. Thus

$$\mathcal{C}_{\mathcal{U}}(u) = \|Q(Au-b)\|^2 \qquad (u \in \mathcal{V}). \tag{44}$$

Existence and uniqueness follow from strong convexity. By Eq. 44, the exact objective agrees on $\mathcal{V}$ with the restriction of the continuously differentiable quadratic

$$F_0(u) \coloneqq \tfrac{1}{2}\|Ou - y\|^2 + \lambda\|Q(Au-b)\|^2.$$

For $u, v \in \mathcal{H}$ and $h \coloneqq v - u$, direct expansion gives

$$F_0(v) = F_0(u) + \langle \nabla F_0(u), h\rangle + \tfrac{1}{2}\|Oh\|^2 + \lambda\|QAh\|^2.$$

By Eq. 20,

$$F_0(v) \geq F_0(u) + \langle \nabla F_0(u), v-u\rangle + \tfrac{\gamma}{2}\|v-u\|^2.$$

Hence F_0, and therefore F on $\mathcal{V}$, is γ-strongly convex. Fixing any $u_0 \in \mathcal{V}$ in the last inequality also shows that $F_0(u) \to \infty$ as $\|u\| \to \infty$ with $u \in \mathcal{V}$. The function is continuous and convex, hence weakly lower semicontinuous, while a closed convex subset of a Hilbert space is weakly closed. The direct method applied to a bounded minimizing sequence therefore gives a minimizer on $\mathcal{V}$, and strong convexity makes the minimizer unique. Denote the minimizer by u_{aff}.

The remaining step derives the deterministic perturbation bound. The variational inequality for minimizing the differentiable convex function F_0 over $\mathcal{V}$ is

$$\langle \nabla F_0(u_{\mathrm{aff}}), v - u_{\mathrm{aff}}\rangle \geq 0 \qquad (v \in \mathcal{V}).$$

Set $h \coloneqq u_{\mathrm{aff}} - u^\star$. Taking $v = u^\star$ gives $\langle \nabla F_0(u_{\mathrm{aff}}), h\rangle \leq 0$. The gradient of F_0 is

$$\nabla F_0(u) = O^*(Ou - y) + 2\lambda A^* Q(Au - b),$$

where $Q^*Q = Q$ was used. The strong monotonicity of ∇F_0, which follows also by subtracting this affine gradient at two points, yields

$$\gamma\|h\|^2 \leq \langle \nabla F_0(u_{\mathrm{aff}}) - \nabla F_0(u^\star), h\rangle \leq -\langle \nabla F_0(u^\star), h\rangle \leq \|\nabla F_0(u^\star)\|\|h\|.$$

If $h = 0$ the desired bound is immediate; otherwise divide by $\gamma\|h\|$. Since $y = Ou^\star + \xi$ and $Au^\star - b = e - J\mu^\star$,

$$\nabla F_0(u^\star) = -O^*\xi + 2\lambda A^* Qe,$$

because $QJ\mu^\star = 0$. The triangle inequality now gives Eq. 21. If $e = \xi = 0$, the right-hand side of Eq. 21 vanishes, which proves the exact-recovery specialization.

Finally, in finite-dimensional $\mathcal{H}$, the quadratic form in Eq. 20 is continuous and vanishes exactly on $\ker O \cap \ker(QA)$. Coercivity clearly implies that this intersection is trivial. Conversely, if the intersection is trivial, the quadratic form is strictly positive on the compact unit sphere and therefore has a positive minimum γ_λ. Because this argument applies to each arbitrary fixed $\lambda > 0$ and the zero set is independent of the value of λ, the resulting constant may be denoted $\gamma_\lambda > 0$ for every positive weight. □

Proof of Proposition 5. Let $\widetilde{F}(u) \coloneqq \|Ou - y\|^2/2 + \lambda\widetilde{\mathcal{C}}(u)$. The assumed uniform profile accuracy implies on $\mathcal{V}$ that

$$F(u) \leq \widetilde{F}(u) \leq F(u) + \lambda\eta_{\mathrm{prof}}.$$

Using first the $\varepsilon_{\mathrm{opt}}$-minimality of $\widetilde{u}$, then testing the infimum with u_{aff}, gives

$$F(\widetilde{u}) \leq \widetilde{F}(\widetilde{u}) \leq \inf_{u\in\mathcal{V}} \widetilde{F}(u) + \varepsilon_{\mathrm{opt}} \leq \widetilde{F}(u_{\mathrm{aff}}) + \varepsilon_{\mathrm{opt}} \leq F(u_{\mathrm{aff}}) + \lambda\eta_{\mathrm{prof}} + \varepsilon_{\mathrm{opt}}.$$

The variational inequality at u_{aff} and γ-strong convexity imply

$$F(\widetilde{u}) \geq F(u_{\mathrm{aff}}) + \tfrac{\gamma}{2}\|\widetilde{u} - u_{\mathrm{aff}}\|^2.$$

Combining the last two inequalities yields

$$\|\widetilde{u} - u_{\mathrm{aff}}\| \leq \sqrt{\frac{2(\varepsilon_{\mathrm{opt}} + \lambda\eta_{\mathrm{prof}})}{\gamma}}.$$

The triangle inequality with Eq. 21 proves Eq. 22. □

E.12 Local Tangent Projection

We first record the projection and nuisance-orthogonality identities used in the tangent analysis and then prove the second-order bias bound. Throughout, for fixed u, we write $J_\mu(u) \coloneqq \partial_\mu R_{\hat\mu}(u)\Sigma^{1/2} : \mathbb{R}^p \to \mathcal{Y}$, whose range $\mathcal{T}_{\hat\mu}(u)$ is finite dimensional, hence closed, so the orthogonal projector $\Pi_{\mathcal{T}_{\hat\mu}(u)}$ is well defined. For $r \coloneqq R_{\hat\mu}(u)$, $J \coloneqq J_\mu(u)$, and $\Pi \coloneqq \Pi_{\mathrm{Range}(J)}$, orthogonal decomposition gives

$$\|r + Jv\|^2 = \|(I-\Pi)r\|^2 + \|\Pi r + Jv\|^2.$$

Since $\Pi r \in \mathrm{Range}(J)$ there exists v with $Jv = -\Pi r$. Thus $\min_v \|r + Jv\|^2 = \|(I-\Pi)r\|^2 = \mathcal{B}_{\hat\mu}(u)$; the approximation in Sec. D.9 concerns only replacing the nonlinear residual family by the linearized residual family.

The same projector gives the nuisance-orthogonality property referred to in Sec. D.9. Let $Q(u) \coloneqq I - \Pi_{\mathrm{Range}(J_\mu(u))}$. For the local residual map

$$\psi(u,\eta) \coloneqq Q(u)(R_{\hat\mu}(u) + J_\mu(u)\eta), \qquad \text{we have} \qquad \partial_\eta \psi(u,\eta)|_{\eta=0} = Q(u)J_\mu(u) = 0.$$

Thus first-order changes in the nuisance physics parameter do not change the estimating residual used to train u.

Lemma 7 (Second-order bias of the tangent-projected residual)**.** *Let $u^\star$ satisfy $R_{\mu^\star}(u^\star) = 0$, let $\hat\mu$ be the nominal parameter, and write $\Delta \coloneqq \mu^\star - \hat\mu$ for the true parameter perturbation. Assume $\mu \mapsto R_\mu(u^\star)$ is twice continuously differentiable on a convex neighborhood of the segment $[\hat\mu, \mu^\star]$ with $\|\partial^2_{\mu\mu} R_\mu(u^\star)\| \le H_\mu$ there, and that the true perturbation lies in the tangent model in the sense that*

$$\partial_\mu R_{\hat\mu}(u^\star)\Delta \in \mathcal{T}_{\hat\mu}(u^\star) = \mathrm{Range}(\partial_\mu R_{\hat\mu}(u^\star)\Sigma^{1/2}),$$

which holds in particular whenever $\Delta \in \mathrm{Range}(\Sigma^{1/2})$, including every case with nonsingular Σ. Then, with $\Pi \coloneqq \Pi_{\mathcal{T}_{\hat\mu}(u^\star)}$,

$$\sqrt{\mathcal{B}_{\hat\mu}(u^\star)} = \|(I-\Pi)R_{\hat\mu}(u^\star)\| \le \tfrac{1}{2}H_\mu\|\Delta\|^2,$$

whereas the unprojected nominal residual satisfies

$$\big|\|R_{\hat\mu}(u^\star)\| - \|\partial_\mu R_{\hat\mu}(u^\star)\Delta\|\big| \le \tfrac{1}{2}H_\mu\|\Delta\|^2.$$

Proof of Lemma 7. Define $G : [0,1] \to \mathcal{Y}$ by $G(s) \coloneqq R_{\hat\mu + s\Delta}(u^\star)$; the segment stays in the neighborhood by convexity, G is twice continuously differentiable, $G(1) = R_{\mu^\star}(u^\star) = 0$, $G'(0) = \partial_\mu R_{\hat\mu}(u^\star)\Delta$, and $G''(s) = \partial^2_{\mu\mu} R_{\hat\mu+s\Delta}(u^\star)[\Delta,\Delta]$ with $\|G''(s)\| \le H_\mu\|\Delta\|^2$. Taylor's theorem with integral remainder gives

$$0 = G(1) = G(0) + G'(0) + \int_0^1 (1-s)G''(s)ds,$$

hence

$$R_{\hat\mu}(u^\star) = -\partial_\mu R_{\hat\mu}(u^\star)\Delta - \rho, \qquad \rho \coloneqq \int_0^1 (1-s)G''(s)ds, \quad \|\rho\| \le \tfrac{1}{2}H_\mu\|\Delta\|^2. \tag{45}$$

Apply $I - \Pi$ to Eq. 45. By the tangent-model assumption, $\partial_\mu R_{\hat\mu}(u^\star)\Delta \in \mathcal{T}_{\hat\mu}(u^\star)$, so $(I - \Pi)\partial_\mu R_{\hat\mu}(u^\star)\Delta = 0$, and since $I - \Pi$ is an orthogonal projector, $\|I - \Pi\| \le 1$. Therefore

$$\|(I-\Pi)R_{\hat\mu}(u^\star)\| = \|(I-\Pi)\rho\| \le \|\rho\| \le \tfrac{1}{2}H_\mu\|\Delta\|^2,$$

and the first bound follows from $\mathcal{B}_{\hat\mu}(u^\star) = \|(I-\Pi)R_{\hat\mu}(u^\star)\|^2$. The second bound is the triangle inequality applied to Eq. 45. □

E.13 Tangent-Profile Stability and Predictor Bias

Proof of Lemma 3. Write $\Pi(u)$ for the orthogonal projector onto $\mathrm{Range}(J(u))$, which is finite dimensional and hence closed. By the projection identity above,

$$b(u) = \|(I - \Pi(u))R(u)\| = \min_{v\in\mathbb{R}^p} \|R(u) + J(u)v\|.$$

The mean value inequality on the convex set $\mathcal{N}$ and the derivative bounds give

$$\|R(u) - R(w)\| \le B_R\|u - w\|, \qquad \|J(u) - J(w)\| \le B_J\|u - w\|, \qquad u, w \in \mathcal{N}, \tag{46}$$

where $\|J(u) - J(w)\|$ denotes the operator norm of a map $\mathbb{R}^p \to \mathcal{Y}$.

The uniform nondegeneracy condition gives $J(u)^*J(u) \succeq \sigma^2 I_p$. Thus $J(u)^*J(u)$ is invertible, $J(u)(J(u)^*J(u))^{-1}J(u)^* = \Pi(u)$, and

$$v(u) \coloneqq -(J(u)^*J(u))^{-1}J(u)^*R(u)$$

is the unique solution of the normal equations, hence the unique profile minimizer, and satisfies

$$J(u)v(u) = -\Pi(u)R(u), \qquad \|R(u) + J(u)v(u)\| = b(u), \qquad \|v(u)\| \le \frac{\|R(u)\|}{\sigma} \le \frac{M_R}{\sigma}.$$

The norm bound follows from $\sigma\|v(u)\| \le \|J(u)v(u)\| = \|\Pi(u)R(u)\| \le \|R(u)\|$. Using $v(w)$ as a candidate in the profile at u, then Eq. 46 and the same norm bound, gives

$$\begin{aligned} b(u) &\le \|R(u) + J(u)v(w)\| \\ &\le \|R(w) + J(w)v(w)\| + \|R(u) - R(w)\| + \|(J(u) - J(w))v(w)\| \\ &\le b(w) + B_T\|u - w\|. \end{aligned}$$

Exchanging u, w yields

$$|b(u) - b(w)| \le B_T\|u - w\|. \tag{47}$$

□

Proof of Theorem 6. Abbreviate $\mathcal{L} \coloneqq \mathcal{L}_{\mathrm{data}}$, $R(u) \coloneqq R_{\hat{\mu}}(u)$, and $J(u) \coloneqq \partial_\mu R_{\hat{\mu}}(u)\Sigma^{1/2}$, and write $\Pi(u)$ for the orthogonal projector onto $\mathrm{Range}(J(u))$. By the projection identity and Lemma 3,

$$b(u) \coloneqq \sqrt{\mathcal{B}_{\hat{\mu}}(u)} = \|(I - \Pi(u))R(u)\| = \min_{v \in \mathbb{R}^p} \|R(u) + J(u)v\|$$

is B_T-Lipschitz, and the unique minimizer $v(u)$ satisfies $\|v(u)\| \le M_R/\sigma$.

Let $u_\lambda \in \mathcal{N}$ be a local minimizer of $\mathcal{J}_{\tan}$ and write $d \coloneqq \|u_\lambda - u^\star\|$. If $d = 0$ the bound holds trivially, so assume $d > 0$.

We first build a smooth majorant and derive the first-order condition of that majorant. Let $v_\lambda \coloneqq v(u_\lambda)$ be the exact profile defined above and set

$$G(u) \coloneqq R(u) + J(u)v_\lambda, \qquad \varphi(u) \coloneqq \mathcal{L}(u) + \lambda\|G(u)\|^2.$$

By the min-representation of $\mathcal{B}_{\hat{\mu}}$, $\|G(u)\|^2 \ge \mathcal{B}_{\hat{\mu}}(u)$ for every $u \in \mathcal{N}$, while $\|G(u_\lambda)\| = b(u_\lambda)$; hence $\varphi \ge \mathcal{J}_{\tan}$ on $\mathcal{N}$ with equality at u_λ. Exactly as in the proof of Lemma 6, local minimality of u_λ for $\mathcal{J}_{\tan}$ therefore transfers to the majorant, so u_λ is a local minimizer of φ. The map G is Fréchet differentiable on $\mathcal{N}$ because the map $u \mapsto J(u)v_\lambda$ is the composition of the differentiable map $u \mapsto J(u)$ with the bounded linear evaluation map $A \mapsto Av_\lambda$ from $L(\mathbb{R}^p, \mathcal{Y})$ to $\mathcal{Y}$, so $DG(u)h = D_uR(u)h + (D_uJ(u)h)v_\lambda$ and, by the tangent-profile regularity assumptions and the profile bound above,

$$\|DG(u)\| \le B_R + B_J\|v_\lambda\| \le B_T, \qquad u \in \mathcal{N}.$$

By the chain-rule identity recorded above, φ is Fréchet differentiable with $\nabla\varphi(u) = \nabla\mathcal{L}(u) + 2\lambda DG(u)^*G(u)$, and the interior first-order condition at the local minimizer gives $\nabla\varphi(u_\lambda) = 0$. Therefore

$$\|\nabla\mathcal{L}(u_\lambda)\| = 2\lambda\|DG(u_\lambda)^*G(u_\lambda)\| \le 2\lambda B_T b(u_\lambda).$$

We next bound the projected residual at the truth. The true-model, curvature, and tangent-coverage assumptions match the hypotheses of Lemma 7, which yields

$$b(u^\star) = \|(I - \Pi(u^\star))R(u^\star)\| \le \tfrac{1}{2}H_\mu\|\Delta\|^2.$$

The pieces now assemble. Strong monotonicity of $\nabla\mathcal{L}$ on the convex set $\mathcal{N}$, derived exactly as in Eq. 32, together with $\nabla\mathcal{L}(u^\star) = 0$ and the Cauchy–Schwarz inequality, gives $md \le \|\nabla\mathcal{L}(u_\lambda)\|$. Combining the strong-monotonicity bound with the majorant bound, Eq. 47, and the residual bound at the truth,

$$md \le 2\lambda B_T b(u_\lambda) \le 2\lambda B_T(b(u^\star) + B_T d) \le \lambda B_T H_\mu\|\Delta\|^2 + 2\lambda B_T^2 d.$$

Since $2B_T^2\lambda < m$, rearranging gives the claimed bound. □

Parameter-affine families. For a parameter-affine family satisfying the tangent-coverage condition of Theorem 6, the map $\mu \mapsto R_\mu(u^\star)$ has $H_\mu = 0$, and Theorem 6 yields exact recovery, $u_\lambda = u^\star$. When the unconstrained affine profile is feasible and the tangent parameterization has full parameter rank, the tangent-profile recovery matches the global profile described in Sec. D.10.

Perturbed data gradients. The perturbation extension of Theorem 2 applies verbatim: If the vanishing data-gradient assumption is weakened to $\|\nabla \mathcal{L}_{\mathrm{data}}(u^\star)\| \le \varepsilon$, then the assembly above gives the same bound with numerator $\varepsilon + \lambda B_T H_\mu \|\Delta\|^2$.

Large nonlinear mismatch. The theorem quantifies the regime claim of Sec. D.9 that is stress-tested in the non-affine benchmark: The guarantee degrades through the curvature term $H_\mu\|\Delta\|^2$ and any loss of locally uniform smoothness or tangent coverage. In the scalar experiment, the tangent coverage is automatic, so the range condition itself does not fail. The large-mismatch degradation that we observe is consistent with the nonlinear Taylor remainder; because the implemented quotient is box-clipped, active clipping can contribute additional approximation error outside the unconstrained interior regime analyzed by the theorem.

E.14 Regularized and Inexact Profiles

Proof of Lemma 4. Continuity on I gives a minimizer $\beta_\star$. Choose a nearest grid point β_g, so $|\beta_g - \beta_\star| \le h/2$. Since $G_K \subset I$,

$$0 \le \mathcal{C}_{G_K}(u) - \mathcal{C}_I(u) \le f(u,\beta_g) - f(u,\beta_\star) \le L_\beta h/2.$$

For the quadratic bound, an endpoint minimizer belongs to G_K and makes the error zero. Otherwise choose an interior minimizer, where $\partial_\beta f(u,\beta_\star) = 0$. Taylor's theorem with the stated second-derivative bound gives $f(u,\beta_g) - f(u,\beta_\star) \le (H_\beta/2)|\beta_g - \beta_\star|^2 \le H_\beta h^2/8$. □

Proof of Proposition 6. Fix u and suppress the dependence on u. The matrix $H_\tau \coloneqq J^*J + \tau I$ is positive definite, with $H_\tau \succeq \tau I$, whether or not J has full rank. Completing the square gives

$$\Phi_\tau(u,\alpha) = \|r\|^2 - \langle J^*r, H_\tau^{-1}J^*r\rangle + \langle H_\tau(\alpha - \alpha_\tau^{\mathrm{free}}), \alpha - \alpha_\tau^{\mathrm{free}}\rangle,$$

with $\alpha_\tau^{\mathrm{free}} = -H_\tau^{-1}J^*r$. The last term is minimized uniquely at zero, proving the free minimizer and value in Eq. 24 and Eq. 25; substituting the minimizer in the residual gives $Q_\tau r$.

On the coefficient ball, continuity gives attainment and the 2τ-strong convexity of $\Phi_\tau(u,\cdot)$ gives uniqueness. Let α_0 be any minimizer of $f(u,\cdot)$ on that ball. Nonnegativity of the ridge term and comparison with this candidate yield

$$\mathcal{C}_{\mathcal{U}}(u) \le \mathcal{C}_{\mathcal{U},\tau}(u) \le f(u,\alpha_0) + \tau\|\alpha_0\|^2 \le \mathcal{C}_{\mathcal{U}}(u) + \tau\rho^2.$$

A feasible free minimizer must also be the constrained minimizer.

For the envelope statement, compactness of the fixed ball, continuity of Φ_τ, and uniqueness imply continuity of $u \mapsto \alpha_{\rho,\tau}(u)$. This profile continuity follows directly by taking a convergent subsequence of any sequence of minimizers with $u_k \to u$: Candidate comparison makes every subsequential limit minimize $\Phi_\tau(u,\cdot)$, and uniqueness identifies that limit. Writing $\alpha(u) = \alpha_{\rho,\tau}(u)$, candidate comparison also gives

$$\begin{aligned}\Phi_\tau(u+h,\alpha(u+h)) - \Phi_\tau(u,\alpha(u+h)) &\le \mathcal{C}_{\mathcal{U},\tau}(u+h) - \mathcal{C}_{\mathcal{U},\tau}(u)\\ &\le \Phi_\tau(u+h,\alpha(u)) - \Phi_\tau(u,\alpha(u)).\end{aligned}$$

The assumed continuous differentiability, uniformly over the compact coefficient ball near u, and $\alpha(u+h) \to \alpha(u)$ make both bounds equal to $D_u f(u,\alpha(u))[h] + o(\|h\|)$. This candidate comparison proves Fréchet differentiability and Eq. 27. Continuity of the envelope gradient proves continuous differentiability. The penalty has no explicit u-derivative because τ is fixed. □

Proof of Proposition 7. Fix u and write $a = \widehat{\alpha}$, $a_\star = \alpha_{\rho,\tau}(u)$, and $\Phi(\alpha) = \Phi_\tau(u,\alpha)$. The normal cone is a closed convex cone in finite dimensions. Hence there is $z \in N_{\mathcal{U}}(a)$ such that $q = \nabla\Phi(a) + z$ has norm $\kappa_{\mathrm{in}}(u)$. Strong convexity and $\langle z, a_\star - a\rangle \le 0$ give

$$\begin{aligned}\Phi(a_\star) &\ge \Phi(a) + \langle \nabla\Phi(a), a_\star - a\rangle + \tau\|a_\star - a\|^2\\ &\ge \Phi(a) + \langle q, a_\star - a\rangle + \tau\|a_\star - a\|^2.\end{aligned}$$

Maximizing $\kappa_{\mathrm{in}} d - \tau d^2$ over $d \geq 0$ therefore yields $\eta_\tau(u) \leq \kappa_{\mathrm{in}}(u)^2/(4\tau)$.

Optimality of $a_\star$ supplies $z_\star = -\nabla\Phi(a_\star) \in N_{\mathcal{U}}(a_\star)$. The gradient of Φ is 2τ-strongly monotone, while the normal cone is monotone. Consequently

$$2\tau\|a - a_\star\|^2 \leq \langle \nabla\Phi(a) - \nabla\Phi(a_\star) + z - z_\star, a - a_\star\rangle = \langle q, a - a_\star\rangle.$$

Cauchy–Schwarz proves the coefficient bound, including the trivial case $a = a_\star$. Feasibility gives $f(u,a) \geq \mathcal{C}_{\mathcal{U}}(u)$, and Eq. 26 gives

$$\begin{aligned} f(u,a) &= \mathcal{C}_{\mathcal{U},\tau}(u) + \eta_\tau(u) - \tau\|a\|^2 \\ &\leq \mathcal{C}_{\mathcal{U}}(u) + \tau\rho^2 + \eta_\tau(u). \end{aligned}$$

Finally, the envelope identity and the assumed Lipschitz property give

$$\|g_{\mathrm{det}} - \nabla_u \mathcal{C}_{\mathcal{U},\tau}(u)\| = \|\nabla_u f(u,a) - \nabla_u f(u,a_\star)\| \leq L_{u\alpha}\|a - a_\star\| \leq \frac{L_{u\alpha}}{2\tau}\kappa_{\mathrm{in}}(u).$$

□

F Experimental Protocol and Reproducibility

F.1 Metrics and Reporting

Data are generated with a true parameter $\mu^\star$, while the physics loss is evaluated using a corrupted nominal parameter $\hat{\mu}$. We report prediction error, true residual, nominal residual, compatibility residual, and two degradation metrics, GDS and the harm threshold. The prediction error and the GDS are

$$\mathrm{RelL2} \coloneqq \frac{\|u_\theta - u^\star\|_2}{\|u^\star\|_2}, \qquad \mathrm{GDS} \coloneqq \frac{1}{N_\delta}\sum_{k=1}^{N_\delta} \frac{\mathrm{Err}_{\mathrm{meth}}(\delta_k)}{\mathrm{Err}_{\mathrm{data}}(\delta_k) + 10^{-12}},$$

where $\delta_1, \ldots, \delta_{N_\delta}$ are the tested mismatch levels and $\mathrm{Err}_{\mathrm{meth}}$ and $\mathrm{Err}_{\mathrm{data}}$ are the seed-averaged RelL2 of the method and of data-only at that level. The score GDS is therefore the mean ratio of a method's seed-averaged RelL2 to the corresponding data-only RelL2 across the mismatch grid. For the multi-trajectory summary, we compute this seed-averaged ratio separately for each trajectory–mismatch pair and then average over those pairs.

Unless stated otherwise, every residual column is an empirical residual MSE,

$$\widehat{\mathrm{Res}}_\mu(u) \coloneqq \frac{1}{n_c}\sum_{i=1}^{n_c} \|r_\mu(u; Z_i)\|_2^2,$$

or the analogous full-grid mean for operator experiments; the tables do not report the square root of the residual MSE. The harm threshold is the smallest tested mismatch ratio δ for which a method's seed-averaged RelL2 exceeds the data-only error by more than 2%. A higher value indicates a later crossing as δ increases; ∞ means that the method stays within the error threshold at all tested ratios. This ordering is by numerical δ, not by distance from the correctly specified value $\delta = 1$. Lower RelL2, residual, and GDS values are better.

The following reporting conventions hold in every table unless the table caption states otherwise. Each entry uses $S = 4$ seeds, except the public compressible-flow check, which uses $S = 2$ data-split seeds. For entries with a $\pm$ term, let $\mathcal{C}$ contain the fixed conditions summarized: the listed mismatch levels and, for multi-trajectory results, trajectory–mismatch pairs. For a metric $x_{s,c}$ from seed s and condition c, we first average within each seed and then report the mean and one standard error across seeds:

$$\bar{x}_s = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} x_{s,c}, \qquad \bar{x} = \frac{1}{S}\sum_{s=1}^{S} \bar{x}_s, \qquad \text{standard error} = \sqrt{\frac{\sum_{s=1}^{S}(\bar{x}_s - \bar{x})^2}{S(S-1)}}.$$

Thus $\pm$ measures uncertainty in the mean across seeds, conditional on the reported conditions. Conditions sharing a seed are kept together, including repeated data-only values across mismatch levels. For a single condition, the reporting formula reduces to the usual standard error over the seeds for that condition.

Paired uncertainty. For a comparison with baseline b, we preserve the same condition–seed pairing and form differences before estimating uncertainty:

$$d_s = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}(x_{s,c}^{\mathrm{compat}} - x_{s,c}^{b}), \quad \bar{d} = \frac{1}{S}\sum_{s=1}^{S} d_s, \quad \mathrm{CI}_{95\%} = \bar{d} \pm t_{0.975,S-1}\sqrt{\frac{\sum_{s=1}^{S}(d_s-\bar{d})^2}{S(S-1)}}.$$

Negative RelL2 differences favor compatibility training. The interval treats seeds as the independent replication units, conditional on the fixed condition grid; mismatch levels and trajectories do not increase S. The CI is a parametric paired t interval, with exact coverage under independent, identically normally distributed seed differences. With $S = 4$, we use $t_{0.975,3} = 3.1824$ and also report all four differences in App. G.1 so the dispersion of the four seed differences is visible without a distributional assumption. Intervals are pointwise, without a multiplicity adjustment, and are not simultaneous guarantees across comparisons. The two-split computational fluid dynamics (CFD) check instead reports individual split differences.

Rows labeled tuned select λ by the MSE on held-out validation observations; the remaining rows use a fixed physics weight. Validation-selected and common-weight comparisons occupy separate tables or explicitly labeled blocks. Boldface identifies the lowest observed mean within a block and does not denote a significance test. App. F.6 gives the resulting replicate counts.

For extrapolation and operator-learning runs, we also record low-dimensional quantities of interest (QoIs). In one-dimensional time-dependent PDEs, we report the mean absolute error (MAE) of the spatial average over evaluation times and the final-time RelL2; for the compressible Navier–Stokes operator, this spatial-average error is computed on the density channel and appears as Mass MAE. In Darcy FNO runs, we report the absolute error of the spatial mean pressure and the boundary MSE. These diagnostics test whether RelL2 gains coincide with physically relevant aggregate predictions rather than only pointwise error reduction.

F.2 Constructing a Family before Evaluation

The family specifies the physical uncertainty available to the learner. The family construction can use measurement tolerances, constitutive restrictions, historical calibration records, or samples from a scientific discrepancy prior. These sources need not reveal the true physics of an evaluation system. A practical construction separates the following three choices.

Structure and sharing. Choose which coefficients, constitutive terms, or spatial modes may vary. Specify whether one nuisance parameter is shared across a trajectory or whether each input function has a separate correction. This choice encodes a physical hypothesis; changing the sharing pattern changes the admissible family.

Basis and extent. Use known correction directions or estimate a small basis from discrepancy records separate from the evaluation systems. Fix the basis normalization before setting the bounds on the basis coefficients. For a positive scalar, $\mathcal{U}(\hat{\mu}) = [\hat{\mu}/\rho, \rho\hat{\mu}]$ with $\hat{\mu} > 0$ and $\rho \geq 1$ expresses a multiplicative tolerance. For example, calibration pairs $(\hat{\mu}_j, \mu_j^{\mathrm{eff}})$ supply log discrepancies $d_j = \log(\mu_j^{\mathrm{eff}}/\hat{\mu}_j)$. A prescribed tolerance or an empirical quantile of $|d_j|$ can set r and hence $\rho = \exp(r)$, with uncertainty in estimated effective coefficients accounted for in the tolerance. For field corrections, express the calibration discrepancies in the chosen parameterization and project the calibration discrepancies onto the fixed normalized basis. An empirical quantile of the resulting coefficient norms can define a candidate radius, while a covariance-shaped bound can retain differences in scale across modes. Compare candidate dimensions and radii using separate validation observations as described next. A radius increase cannot add a missing direction.

Selection and freezing. If several constructions are compared, select a construction using held-out observation error with a declared search budget, then freeze the rule before final evaluation. Validation observations can assess prediction without revealing physical coefficients. The minimum profiled residual alone is unsuitable for comparing nested families because the minimum profiled residual can only decrease as the family grows. Freeze the basis, normalization, bounds, and sharing pattern; the profiled coefficients still change with the predictor during training. The fixed rule may

depend on a new system's observed nominal inputs, but not on that system's held-out test solution or true correction.

The tolerance and validation recipes above are application procedures, not claims that the reported radii were calibrated using those recipes. An empirical quantile alone supplies no guarantee of true-model containment on a new system. After a constructed family is fixed, the deterministic recovery results apply under the hypotheses stated for each result; Corollary 2 describes the remaining approximation error when no member is exactly compatible. If family construction is independent of the collocation sample and the collocation points are independent and identically distributed, Theorem 3 can also be applied conditional on that family, with the constants and covering number that correspond to that family. Neither statement estimates the probability of containment from calibration data.

What the reported families test. The scalar and reaction-law families below are controlled constructions: Nominal parameters are corrupted from known generating values, and radius sweeps vary containment within that design. The prescribed field-Darcy basis is shared with the corruption generator, exposing the effect of matching the adjustment directions. The learned-basis check instead estimates the correction modes from a separate synthetic discrepancy ensemble, with a fixed coefficient radius. These settings respectively test the extent, structure, and independent construction of a family. The reported coefficient-distribution shift keeps the discrepancy family fixed and tests a different source of generalization.

F.3 Synthetic and Public PDE Benchmarks

PINN benchmarks. We use periodic viscous Burgers,

$$u_t + uu_x = \nu u_{xx}, \qquad \nu^\star = 0.02,$$

and periodic Fisher–KPP reaction–diffusion,

$$u_t = Du_{xx} + \alpha u(1-u), \qquad D^\star = 0.01, \quad \alpha^\star = 1.0.$$

For synthetic Burgers, training uses $\hat{\nu} := \delta\nu^\star$ and the multiplicative interval $[\hat{\nu}/\rho, \rho\hat{\nu}]$. For reaction–diffusion, only the nominal reaction rate is corrupted, $\hat{D} := D^\star$ and $\hat{\alpha} := \delta\alpha^\star$; the profiled and inverse families use $D \in [D^\star/2, 2D^\star]$ and $\alpha \in [\hat{\alpha}/\rho, \rho\hat{\alpha}]$.

Non-affine reaction-law benchmark. To test non-affine uncertainty, we also use

$$u_t = Du_{xx} + \alpha \exp(\beta u)u(1-u), \qquad D^\star = 0.01, \quad \alpha^\star = 1.0, \quad \beta^\star = 0.$$

Training corrupts the reaction law by setting $\hat{\beta} := \log\delta$ for mismatch ratio δ and $\mathcal{U} := [\beta_{\min}, \beta_{\max}] = [\hat{\beta} - \log\rho, \hat{\beta} + \log\rho]$. The implemented global compatibility loss profiles β on a 25-point deterministic grid, while the implemented tangent loss linearizes the residual in β and adds the clipped least-squares quotient of that linearization. App. F.6 gives the two formulas, the numerical safeguards, and the regime in which the implemented quotient coincides with the projector analyzed by Theorem 6.

Neural-field training setup. Solutions are generated by a pseudo-spectral fourth-order Runge–Kutta scheme on a 256×101 grid (Trefethen, 2000). All PINN methods use the same Fourier-feature coordinate multilayer perceptron (MLP) (Tancik et al., 2020) with 64 Fourier features, four hidden layers, width 128, sigmoid linear unit activations, AdamW (Loshchilov & Hutter, 2019), 2500 steps, 4096 interior observations plus the full initial condition, and 4096 collocation points. Mismatch ratios are $\{0.25, 1, 4, 16\}$.

For the time-extrapolation stress test, we train reaction–diffusion models only from observations with $t \leq 0.5$ and evaluate RelL2 and residuals only on $t \geq 0.5$. This test uses mismatch 16, 512 interior observations in the early window, the full initial condition, 4096 collocation points, and the same architecture and optimizer.

PDEBench trajectory checks. Alongside the synthetic trajectories, we evaluate standardized PDEBench data (Takamoto et al., 2022). We use the public $\nu = 0.02$ Burgers dataset, take a

single trajectory from that dataset, and downsample that trajectory to a 101×256 space-time grid. The true viscosity is $\nu^\star = 0.02$ and the nominal viscosity is corrupted by the same mismatch ratios $\{0.25, 1, 4, 16\}$. The coordinate MLP, optimizer, 2500-step schedule, full initial condition, and 4096 collocation points match the synthetic Burgers runs. The loaded trajectory spans $t \in [0, 2]$; the 1024 interior observations, held-out validation points, and collocation points are restricted to $t \leq 1$, while every reported trajectory metric is evaluated on the full interval $t \in [0, 2]$. Thus this check combines observed-window fitting with second-half temporal extrapolation rather than full-domain interpolation.

A multi-trajectory PDEBench check on public trajectory indices $\{0, 1, 3\}$ uses the same 2500-step recipe, the same $t \leq 1$ training, validation, and collocation window, full $t \in [0, 2]$ evaluation, and the two representative mismatch ratios $\{1, 16\}$. The primary multi-trajectory comparison fixes $\lambda = 0.2$ for every physics method; additional validation-selected nominal and inverse results are reported separately.

Compressible flow with an auxiliary momentum loss. The compressible-flow check evaluates an auxiliary momentum loss with uncertain viscosity. In the equation used to generate the data, viscous forces depend on the fluid density and spatial density variation. Our auxiliary loss computes the viscous term as if density were fixed at one, so even the true solution can yield a nonzero auxiliary residual.

A multi-variable public PDEBench operator family uses the one-dimensional compressible Navier–Stokes benchmark with random periodic initial conditions and shear and bulk viscosities $\eta = \zeta = 0.1$. The FNO maps three initial-condition fields, density, velocity, and pressure, to the full three-channel trajectory. We channel-normalize the supervised target, use the same width-64, 16-mode, 200-epoch FNO recipe and split of 512 training and 128 test trajectories, and downsample to a 26×256 space-time grid. The data generator uses kinematic shear and bulk coefficients, with $\kappa := \zeta + 4\eta/3$, and the density-weighted one-dimensional viscous momentum residual of the generator is

$$\widetilde{R}_\kappa(\rho, v, p) := \rho(v_t + vv_x) + p_x - \partial_x(\kappa\rho v_x).$$

All reported CFD table values use the simpler unit-density auxiliary diagnostic

$$R^{\mathrm{sur}}_\kappa(\rho, v, p) := \rho(v_t + vv_x) + p_x - \kappa v_{xx}.$$

This surrogate is an uncertain auxiliary residual. For each current prediction, compatibility training selects the allowed viscosity coefficient κ that minimizes the squared norm of this auxiliary residual and adds that minimum to the data loss with weight λ. For constant κ, the two continuum residuals satisfy

$$R^{\mathrm{sur}}_\kappa - \widetilde{R}_\kappa = \kappa[(\rho - 1)v_{xx} + \rho_x v_x].$$

Consequently, even at the generator coefficient $\kappa^\star$, the surrogate need not vanish on the true trajectory. Varying one scalar coefficient adjusts only the v_{xx} direction and generally cannot absorb this entire structural discrepancy. The experiment therefore evaluates the usefulness of an uncertain auxiliary loss. Surrogate-residual reduction measures agreement with that loss, while RelL2 and the density-mass QoI assess prediction; the auxiliary-loss experiment does not establish exact recovery of the generating PDE.

The generator coefficient is $\kappa^\star = 0.2333$, and training corrupts that value by mismatch ratios $\{1, 16\}$. Every reported physics-informed method uses the common weight $\lambda = 10^{-4}$; the adjacent values $\{2 \cdot 10^{-5}, 5 \cdot 10^{-4}\}$ are sensitivity runs only.

F.4 Baselines and Weight Selection

Baselines. We compare data-only, nominal PINN, inverse PINN with bounded learnable parameters, worst-case residual, uncertainty-average residual, compatibility training, and tangent-projected training. These experiments examine residual objectives and uncertainty-family design; learned discrepancy-network methods are not included in this comparison. The uncertainty-average baseline minimizes the average residual over a deterministic grid in $\mathcal{U}$ rather than the compatibility residual. The worst-case residual is a robust baseline, used with the common fixed physics weight in the affine and Darcy sweeps and as a tuned baseline in the non-affine sweep. For the non-affine reaction-law benchmark, the worst-case baseline maximizes over the same deterministic β grid used by global profiling, not only over interval endpoints.

Joint updates and empirical profiles. Inverse PINNs jointly update bounded physical parameters and the predictor; compatibility training recomputes the physics-parameter profile before each predictor update. The inverse-PINN comparisons therefore test how these two update procedures affect prediction within the reported training budgets. Proposition 2 shows that exact joint optimization and profiling have the same globally optimal predictors when the predictor class, loss, physics-parameter family, and sharing rule are the same. For these matched comparisons, the observed gaps reflect the implemented training procedures and numerical approximations.

The inverse baselines use a persistent bounded parameter initialized at the nominal value. Each component is represented as $\mu = \mu_{\min} + (\mu_{\max} - \mu_{\min})\,\mathrm{sigmoid}(z)$, and the corresponding trainable logit z is included with the predictor parameters in AdamW, using the same learning-rate schedule and weight decay. Compatibility training instead recomputes the nuisance profile before each predictor update. The parameterizations handle boundaries differently: In exact arithmetic, finite logits reach only the interval interior, whereas clipped and grid profiles can attain endpoints. At mismatch and radius 16, the true Burgers viscosity and reaction coefficient lie at the respective parameter lower bounds, as does $\beta^\star = 0$ in the non-affine interval. Finite-budget inverse behavior can therefore reflect boundary approach and parameter scaling as well as predictor–nuisance coupling. The regular interior Hessian result of Proposition 4 does not directly describe those boundary cases.

For neural-field training, each iteration draws a fresh collocation batch and profiles one parameter vector across that batch. Scalar Darcy and compressible-flow profiles similarly share one scalar across the current operator minibatch, whereas the field-valued Darcy solve profiles a separate coefficient vector for each input function. The inverse scalar is shared across all training iterations. Thus per-batch profiling is an empirical approximation to a shared population profile, with the distinction explained in Sec. D.5. The non-affine 25-point grid and the stabilized affine and field solves introduce the additional approximations specified in Sec. D.10 and below.

The reported inverse comparisons consequently measure the performance of these finite-budget training procedures. Proposition 2 characterizes matched exact objectives; the empirical differences do not establish distinct global predictor optima or identify conditioning as the sole cause of a performance gap.

Physics-weight selection. Table 6 specifies the selection rule for each comparison. In the affine, non-affine, and PDEBench validation-selected blocks, every physics method uses $\Lambda_4 = \{0.02, 0.05, 0.2, 1.0\}$, the same held-out observations, and the MSE on those observations as the selection metric. Selection is performed separately for each mismatch–seed cell and, for multiple trajectories, each trajectory; the selected run is then evaluated with the reported prediction and residual metrics. The neural-field runs share the architecture, observations, collocation budget, and 2500 predictor-update steps within each setting. Equal update counts do not imply equal wall-clock cost. In the main synthetic affine and non-affine sweeps, nominal, inverse, and compatibility training each select $\lambda = 0.02$ in all 16 mismatch–seed cells per benchmark. Thus the selected candidate is constant across these cells; this observation does not separate the variance of selection from other seed-dependent variation.

Common-weight blocks measure performance at a shared λ; validation-selected blocks compare methods under a matched search budget. The affine worst-case and non-affine tangent rows have no complete validation grid, so the corresponding procedures are compared with other methods at the original fixed weights, 0.2 and 0.05, respectively. The multi-trajectory, extrapolation, sparse, and noisy comparisons use the common fixed weight 0.2. These blocks are assembled from existing runs at those weights, without selecting the weight by test performance. Additional validation-selected nominal and inverse results are retained in separate blocks: The multi-trajectory check uses Λ_4, and the extrapolation checks use $\{0.02, 0.05, 0.1, 0.2\}$. Compatibility training has only the fixed-weight run in those checks. The uncertainty radius is 16 in the main and trajectory sweeps.

All methods use a scalar physics weight without adaptive scheduling. Validation locations are disjoint from supervised training observations and remain in the observed time window in the trajectory and extrapolation checks. No test error is used for weight selection. The operator comparisons have no independent validation partition and retain the original common fixed weights: $\lambda = 10^{-5}$ for Darcy and $\lambda = 10^{-4}$ for CFD. Neighboring CFD values $\{2 \cdot 10^{-5}, 5 \cdot 10^{-4}\}$ are sensitivity runs for nominal, inverse, and worst-case training. The operator records duplicate test MSE in the `val_mse`

Table 6: Weight-selection protocols. Candidate counts are per physics method and per mismatch–seed setting, and also per trajectory when applicable. The data-only reference has no physics weight. Counts describe the selection rule in each reported block; additional sensitivity runs do not change that rule. N, I, W, A, C, and T denote nominal, inverse, worst-case, uncertainty-average, compatibility training, and tangent, respectively.

Comparison	Physics methods	Selection rule	Trials	Seeds
Affine main	N, I, A, C	Validation MSE, Λ_4	4	4
Non-affine main	N, I, W, A, C	Validation MSE, Λ_4	4	4
PDEBench (single)	N, I, C	Validation MSE, Λ_4	4	4
Affine fixed	N, I, W, A, C	Fixed $\lambda = 0.2$	1	4
Non-affine fixed	N, I, W, A, T, C	Fixed $\lambda = 0.05$	1	4
PDEBench (multiple)	N, I, C	Fixed $\lambda = 0.2$	1	4
Extrapolation (both)	N, I, C	Fixed $\lambda = 0.2$	1	4
Sparse and noisy grid	N, I, C	Fixed $\lambda = 0.2$	1	4
Darcy operators	As listed in each table	Fixed $\lambda = 10^{-5}$	1	4
Public CFD operator	N, I, W, A, C	Fixed $\lambda = 10^{-4}$	1	2

field; this field and any automatically generated tuned rows derived from that field are excluded from model selection and from the reported comparisons.

F.5 Darcy Operator Benchmarks

We also evaluate an operator-learning benchmark following the standard Darcy flow setup used in neural-operator work. Inputs are random log-normal permeability fields $a(x, y)$ and targets solve

$$-\nabla \cdot (a \nabla u) = 1, \qquad u|_{\partial\Omega} = 0.$$

The model receives a corrupted coefficient field $a_{\mathrm{nom}} := \delta a$ and is trained with a residual computed from a_{nom}. The scalar-uncertainty experiment uses 64 training functions, 128 test functions, width 64, 16 modes, 200 epochs, mismatch ratios $\{4, 16\}$, a 32×32 grid, and the common fixed weight $\lambda = 10^{-5}$, with uncertainty radius $\rho = 64$.

We further test field-valued coefficient uncertainty. For each input function, we sample a low-frequency multiplier

$$q_i^{\mathrm{data}}(x, y) := \mathrm{clip}_{[0.25,4]}(1 + \sum_{k=1}^{8} \alpha_{ik} \phi_k(x, y))$$

and give the model the nominal coefficient $a_{\mathrm{nom},i} := a_i / q_i^{\mathrm{data}}$ while the target solves the PDE with a_i. The eight basis functions ϕ_k are fixed low-frequency modes, so the declared family is a truncated spectral expansion of a random coefficient field. Scalar compatibility training profiles only one global scale.

The field row uses the corresponding un-clipped linear multiplier family and a ridge-stabilized approximate inner solve. If $X_i(u)$ is the finite-difference residual design matrix, $y_i(u)$ the target vector, n_i the number of residual entries, and $\tau = 10^{-4}$, the implemented coefficients are

$$\widetilde{\alpha}_i(u) := \left(\frac{X_i(u)^* X_i(u)}{n_i} + \tau I \right)^{-1} \frac{X_i(u)^* y_i(u)}{n_i},$$

$$\alpha_{\mathrm{prof},i}(u) := \min \left\{ 1, \frac{\rho}{\|\widetilde{\alpha}_i(u)\|_2} \right\} \widetilde{\alpha}_i(u).$$

The second line uses the convention $\alpha_{\mathrm{prof},i} := 0$ when $\widetilde{\alpha}_i = 0$. Training and the reported compatibility diagnostic use residual MSE at this detached coefficient vector, without adding the ridge penalty. When the unclipped ridge solution is feasible, Eq. 27 gives the exact regularized-envelope interpretation of the update, up to numerical solve error. Active radial clipping is an approximate constrained solve, with the resulting value and gradient errors characterized in Proposition 7. These coefficient solves provide stabilized profiles of the declared linear family; positivity clipping in the

data generator can also create an out-of-family component. The targets are the same elliptic Darcy solves used in the neural-operator benchmark; only the encoded coefficient field is spatially misspecified.

A stronger field-valued Darcy FNO check raises the supervised baseline with 1024 training functions, 128 test functions, FNO width 64, 16 Fourier modes, batch size 32, and 200 epochs. This long-run setting uses the same field-valued uncertainty family, but averages over only the two harder mismatch amplitudes $\{0.5, 1.0\}$.

As a separate high-resolution check, we rerun the field-valued coefficient benchmark on a 64×64 grid with 512 training functions, 64 test functions, width 64, 16 Fourier modes, batch size 32, and 100 epochs. This setting uses the same eight-dimensional sample-specific coefficient uncertainty and mismatch amplitudes $\{0.5, 1.0\}$. This check tests whether the compatibility residual remains useful when the finite-difference physics diagnostic is evaluated on a finer grid than the main operator sweep.

We additionally test coefficient-distribution shift without changing the grid. The FNO is trained on smoother Darcy coefficients sampled with four Fourier modes and evaluated on independently generated rougher coefficients sampled with ten modes. This out-of-distribution field-valued Darcy check uses 1024 training functions, 128 test functions, width 64, 16 modes, batch size 32, and 150 epochs. The coefficient-error family remains the same eight-dimensional sample-specific field uncertainty, so the experiment isolates operator generalization under a shifted coefficient distribution from out-of-family residual mismatch.

A basis learned from independent discrepancy samples. The learned-basis Darcy check constructs the correction modes from 256 independent samples of a smooth discrepancy prior, separately for each seed. Each prior field z is a sum of sine and cosine modes with frequency indices $k_x, k_y \in \{1, \dots, 10\}$ and independent Gaussian amplitudes scaled by $(k_x^2 + k_y^2)^{-1.15}$, then centered spatially and normalized to unit spatial standard deviation. PCA is applied to the centered matrix of flattened prior samples. We retain eight modes, center each spatially, and normalize each to unit maximum absolute value. The resulting basis is fixed throughout predictor training and shared across the two mismatch amplitudes within a seed. The basis construction uses neither test solutions nor individual test correction fields.

Separate prior draws generate the actual coefficient multipliers,

$$q_i^{\mathrm{data}}(x, y) = \mathrm{clip}_{[0.25,4]}(\exp(0.35\delta z_i(x, y))), \qquad \delta \in \{0.5, 1.0\}.$$

As in the prescribed-basis experiment, $a_{\mathrm{nom},i} = a_i / q_i^{\mathrm{data}}$ is the input and the target solves with a_i. Basis samples and coefficient–solution pairs use separate pseudorandom streams, initialized with seed offsets 86017 and 12853, respectively. The profile uses the linear multiplier $1 + \sum_{k=1}^{8} \alpha_{ik}\phi_k$, with the same ridge solve, radial clipping, and detached updates described above. The nonlinear multiplier need not belong to this linear family. Projections of the true correction onto the basis are computed only for diagnostics and do not enter training.

The experiment fixes the basis dimension at eight, the coefficient radius at $\rho = 1$, and the ridge at $\tau = 10^{-4}$; the prior samples determine the basis directions. The experiment uses 512 training and 64 test functions, a 32×32 grid, FNO width 48, 12 Fourier modes, batch size 32, 80 epochs, and the common physics weight $\lambda = 10^{-5}$. Data-only, nominal, uncertainty-average, and field compatibility training over two mismatch amplitudes and four seeds yields 32 runs. Thus this check evaluates a basis learned independently of the test instances under a known synthetic prior; the learned-basis check does not calibrate a radius or learn the discrepancy prior from physical measurements.

F.6 Implementation and Reproducibility

Results are recorded at the run, tuned-run, and summary levels, and every reported value is computed from the run-level records. Runs are keyed by the full experimental metadata, including the time-window controls of the extrapolation experiments and the basis, grid, and coefficient-roughness metadata of the operator benchmarks, so that summaries are grouped by the actual experimental setting.

The remaining subsections document replicate counts and the seed protocol, the tangent implementation, the execution environment, and consistency audits.

Seeds and replicate counts. Seeds control both the observation draw and the model initialization; for the public operator check, the seeds additionally control the data split.

All synthetic PINN experiments use four seeds per configuration. These experiments comprise the main affine and non-affine mismatch sweeps, the two time-extrapolation checks, the sparse and noisy phase grid, and the scalar radius sweep. The field-valued radius sweep and the synthetic Darcy FNO benchmarks also use four seeds per configuration. The Darcy benchmarks comprise the data-limited scalar run, the prescribed and learned field-basis runs, the stronger long-run check, and the high-resolution and out-of-distribution variants. Thus an aggregate over four mismatch levels contains 16 condition–seed cells, a two-mismatch Darcy aggregate contains eight, a three-mismatch field-valued Darcy aggregate contains 12, and each individual radius cell and each sparse or noisy cell contains four.

The public PDEBench Burgers trajectory checks use four observation and model seeds per trajectory, so the three-trajectory check over two mismatch ratios contains 24 cells, and the tuned baselines of that check select λ within each trajectory–mismatch–seed cell rather than pooling across seeds. The public PDEBench compressible Navier–Stokes operator check uses two seed-indexed contiguous trajectory blocks, with the same training and test partitions shared across methods within each split. Four-seed experiments use $\{0, 1, 2, 3\}$, and the public operator check uses $\{0, 1\}$. Each aggregate has four seed-level replicates, or two for the compressible-flow check, regardless of the number of conditions averaged within each seed. Every run records a seed, so a summary row is reproducible from the run-level records alone.

Sources of variation. The seed varies observation sampling or the data split together with predictor initialization. For tuned rows, the reported seed variation covers the full training-and-selection procedure; for the learned-basis Darcy check, the reported seed variation also includes the independent prior sample used to construct the basis, which is then fixed within each seed. These sources are not crossed in independent repetitions, so the reported standard errors do not separately estimate initialization, sampling, basis-construction, or selection variance.

Non-affine tangent implementation. In the non-affine reaction-law benchmark of Sec. 5.1, the implementation clips the exponent βu to $[-12, 12]$ as an overflow safeguard. Where that safeguard is inactive, the implemented tangent loss uses the exact local linearization

$$R_{\hat{\beta}+\eta}(u) \approx R_{\hat{\beta}}(u) + \eta \partial_\beta R_{\hat{\beta}}(u), \qquad \partial_\beta R_{\hat{\beta}}(u) = -\alpha e^{\hat{\beta} u} u^2 (1 - u).$$

Writing $J_\beta(u) \coloneqq \partial_\beta R_{\hat{\beta}}(u)$ and using the empirical collocation inner product, the scalar quotient is

$$\widetilde{\eta}(u) \coloneqq -\frac{\langle R_{\hat{\beta}}(u), J_\beta(u) \rangle}{\max\{\|J_\beta(u)\|^2, 10^{-12}\}}, \qquad \eta_{\tan}(u) \coloneqq \mathrm{clip}_{[\beta_{\min}-\hat{\beta}, \beta_{\max}-\hat{\beta}]}\, \widetilde{\eta}(u),$$

and the training residual is $R_{\hat{\beta}}(u) + \eta_{\tan}(u) J_\beta(u)$. Theorem 6 analyzes the unconstrained interior projector; the experimental quotient above agrees with that projector whenever the parameter clipping and exponent safeguard are inactive and $\|J_\beta(u)\|^2 > 10^{-12}$. Outside that regime, the quotient is a stabilized tangent surrogate rather than the exact projector analyzed by the theorem.

Execution. The experiments were parallelized across four H100 graphics processing units (GPUs) by assigning independent seeds and sweep configurations to separate devices. Sweeps are resumable and resume-keyed by the full experimental metadata, so a partially completed sweep can be extended without recomputation and without silently regrouping runs. Downsampled PDEBench trajectories and operator subsets are cached and reused across runs.

Consistency checks. Three automated checks run over the collected run records. The first verifies the scalar values and orderings reported in the tables and confirms that every run of the critical sweep grids is present. The second audits the experimental design: Validation-tuned PINN baselines and public-operator sensitivity runs must have complete λ grids, the sparse and noisy phase grid must contain every method, observation count, noise level, and seed cell, and the stronger Darcy FNO

Table 7: Mean recorded seconds per run at matched fixed weights. PINN times cover the training loop; FNO times include the final evaluation. The ratio divides the mean time for compatibility training by the mean time for nominal training. Each PINN row averages 16 seed–mismatch cells, each Darcy row eight, and CFD four. The field row uses the strong eight-mode setting of Table 4; no matched field-inverse run is available.

Setting	Data-only	Nominal	Inverse	Compatibility training	Ratio
Burgers PINN	3.59	17.82	17.94	17.51	0.98
Reaction–diffusion PINN	3.42	18.47	19.63	20.47	1.11
Non-affine PINN	3.52	17.96	18.10	22.22	1.24
Scalar Darcy FNO, 64 inputs	5.64	5.62	5.56	5.53	0.98
Field Darcy FNO, 1024 inputs	42.31	46.10	—	77.64	1.68
PDEBench CFD FNO	51.99	56.54	58.29	59.89	1.06

runs must use the stated training budgets. For the public-operator runs, this audit evaluates the fixed center-weight rows that the tables report, not rows derived from the duplicated validation metric. The second check also rejects duplicate run keys and non-finite or negative core metrics, so partial merges cannot enter the tables.

The third check recomputes the profile summaries discussed in Sec. 5: the non-affine reaction-law profile, the reaction–diffusion profile error relative to inverse PINN, and the field-valued Darcy profile error relative to post-hoc profiling of data-only predictions. The third check then verifies that the numerical values and comparisons in the manuscript agree with the run records. None of these checks is used to select models; the three checks guard against regrouping errors, missing robustness runs, and reported values that disagree with the underlying runs. An additional statistical audit, `exp/statistical_reporting.py`, reconstructs paired comparisons directly from raw run rows, checks matching metadata and complete seed–condition grids, and verifies the paired table entries, weight-selection frequencies, and sparse-grid harm counts. Summary standard deviations and standard errors are computed across seed averages in `exp/results.py`, consistently with the manuscript.

F.7 Computational Cost and Numerical Accuracy

Recorded execution times. Table 7 compares existing run-level records at the common weights used in the fixed-weight prediction tables. Within each setting, methods share the predictor architecture, training schedule, data size, and seed–mismatch cells. We average mismatch levels within each seed and then average seeds, as for the prediction metrics. Ratios divide these mean times; validation-selected rows and the cost of searching multiple weights are not included. This timing protocol separates the cost of adding physics to data-only training from the further cost of profiling rather than using nominal or jointly updated coefficients.

The weights are $\lambda = 0.2$ for both affine PINNs, 0.05 for the non-affine PINN, 10^{-5} for Darcy, and 10^{-4} for CFD. Each PINN run has 2500 predictor updates, giving mean times for compatibility training of 7.00, 8.19, and 8.89 ms per update in table order. The scalar Darcy, field Darcy, and CFD runs each use 200 epochs, with 800, 6400, and 6400 total minibatch updates, respectively. Dividing the recorded total times for the scalar Darcy, field Darcy, and CFD runs by the respective update counts gives 6.91, 12.13, and 9.36 ms per update, with final evaluation amortized into these operator values. The same divisors apply to the other methods in each row, so the relative costs are unchanged.

The timers exclude dataset construction and model setup. PINN times include the training-loop diagnostics, with the same gradient-conflict logging interval of 500 updates for the physics methods in these rows. These timing measurements are elapsed-time records from the H100 execution described above, rather than isolated inner-solver measurements or a dedicated benchmark with synchronized GPU timing boundaries. Small differences near a ratio of one should be read as similar recorded cost. In particular, the $4.88\times$ to $6.31\times$ PINN cost relative to data-only also includes the PDE derivative and physics backpropagation work shared with nominal and inverse training. The stronger field profile has a larger cost, $1.68\times$ nominal and $1.83\times$ data-only, for sample-specific spa-

tial corrections. Predictions of u from the trained network do not require optimizing the physics parameters; the profile is used for training and residual diagnostics.

Dependence on nuisance dimension. Let n be the number of residual entries per profile, p the nuisance dimension, and b the number of independently profiled inputs in a minibatch. The following costs are additional to prediction and construction of the shared PDE derivatives. A scalar affine profile requires $O(n)$ reductions. The two-coefficient reaction–diffusion implementation enumerates $3^2 = 9$ free, lower, and upper active-set configurations and solves the feasible free subproblems. The general active-set enumeration has 3^p configurations and is therefore intended for small p. A scalar grid profile compares K residuals using $O(Kn)$ arithmetic; the non-affine experiment fixes $K = 25$ and reuses the state and the corresponding PDE derivatives across candidates.

For the field-valued Darcy implementation, finite-difference assembly of the p correction columns costs $O(bnp)$. Forming the dense normalized Gram matrices costs $O(bnp^2)$, and the b systems of size p cost $O(bp^3)$. Radial clipping costs $O(bp)$. Although the coefficients are detached, the selected residual still differentiates through the base and correction columns, so the field cost includes the construction of both column types and the backward pass through both. These counts explain why the scalar and field profiles have different costs; the operation counts do not assume that a larger p leaves the residual assembly unchanged.

Memory and approximation controls. The field design tensors and Gram matrices store $O(bnp + bp^2)$ numbers in addition to the predictor's computation. At $b = 32$, $n = 900$, and $p = 8$ in the strong field setting, one 32-bit floating-point design tensor occupies 0.879 MiB and one batch of Gram matrices occupies 8 KiB. These memory estimates are tensor-size counts; autograd intermediates, temporary buffers, and allocator behavior also affect peak device memory. The implementation keeps no differentiation graph through the coefficient solve, while retaining the graph for the selected physics residual.

The tested inner computations are direct affine solves, finite grids, or a direct ridge solve followed by radial clipping. These direct solves and finite-grid evaluations have no iterative inner stopping tolerance. Grid resolution and coefficient regularization therefore specify the implemented accuracy choices. Lemma 4 relates candidate spacing to value error, and Propositions 6 and 7 separate ridge shrinkage from inner optimality error. The records contain no peak-memory measurements or ridge or tolerance sweeps; the memory and accuracy statements here are analytical, while Table 7 reports the existing execution times.

G Additional Results

This section reports additional benchmark tables and diagnostic analyses summarized in Sec. 5.

G.1 Paired Comparisons across Seeds

Tables 8 and 9 quantify prediction differences for the main comparisons using the existing matched runs. Each difference averages the conditions within one seed before the four seeds are summarized. This aggregation within each seed keeps the number of replications separate from the number of conditions covered by a comparison. Strong field Darcy and learned-basis Darcy improve over data-only in every matched cell. The comparison of the learned basis with nominal training varies by seed. For single-trajectory PDEBench and observed-window collocation, the intervals span zero despite favorable mean differences. These intervals describe uncertainty in the mean difference under the reported protocol; an interval spanning zero does not establish equivalence.

G.2 Public PDEBench Validation

Burgers trajectories. Table 10 provides a standardized-data, first-half-to-full-trajectory check of the same failure mode. When the nominal viscosity is correct or close, the tuned nominal and inverse baselines improve the residual and can improve RelL2. Under the largest mismatch, however, tuned nominal and inverse PINNs degrade to RelL2 0.4059 and 0.3615, while compatibility training remains at 0.2747. Across the full mismatch sweep, compatibility training achieves the lowest

Table 8: Paired RelL2 differences: compatibility training minus the named baseline. V denotes per-cell validation selection for both physics methods; F denotes the common fixed weight: 0.05 for non-affine, 10^{-5} for Darcy, and 0.2 otherwise. All intervals use four seed averages. The last two columns count negative seed-average differences and negative condition–seed differences, respectively; these counts are descriptive, not additional independent trials. Future and observed collocation cover the full domain and the observed window, respectively, in the two extrapolation checks.

Label	Setting	Baseline	$\bar{d}$	Pointwise 95% CI	Seeds	Cells
A	Burgers	inverse (V)	-0.00527	$[-0.00573, -0.00481]$	4/4	12/16
B	Reaction–diffusion	inverse (V)	-0.02130	$[-0.02209, -0.02050]$	4/4	9/16
C	Non-affine	inverse (V)	-0.00466	$[-0.00484, -0.00448]$	4/4	15/16
D	Non-affine	tangent (F)	-0.00430	$[-0.00446, -0.00414]$	4/4	14/16
E	PDEBench single	inverse (V)	-0.01429	$[-0.03769, +0.00910]$	4/4	9/16
F	PDEBench multi	inverse (F)	-0.01713	$[-0.03256, -0.00170]$	4/4	10/24
G	PDEBench multi	data-only (F)	-0.04913	$[-0.07042, -0.02784]$	4/4	24/24
H	Strong Darcy	data-only (F)	-0.00114	$[-0.00151, -0.00078]$	4/4	8/8
I	Strong Darcy	scalar profile (F)	-0.00416	$[-0.00456, -0.00376]$	4/4	8/8
J	Learned-basis Darcy	data-only (F)	-0.00296	$[-0.00458, -0.00134]$	4/4	8/8
K	Learned-basis Darcy	nominal (F)	$+0.00036$	$[-0.00060, +0.00132]$	2/4	2/8
L	Future collocation	inverse (F)	-0.36903	$[-0.39308, -0.34498]$	4/4	4/4
M	Observed collocation	inverse (F)	-0.02817	$[-0.07109, +0.01476]$	3/4	3/4

Table 9: Individual seed-average RelL2 differences for the comparisons in Table 8. Each column is one paired replicate; negative values favor compatibility training. Unrounded run records determine the means and intervals.

Label	Seed 0	Seed 1	Seed 2	Seed 3
A	-0.00511	-0.00552	-0.00494	-0.00551
B	-0.02199	-0.02121	-0.02080	-0.02118
C	-0.00471	-0.00467	-0.00449	-0.00475
D	-0.00424	-0.00442	-0.00420	-0.00435
E	-0.03348	-0.00143	-0.00425	-0.01801
F	-0.02744	-0.00403	-0.01812	-0.01893
G	-0.03469	-0.05563	-0.04181	-0.06439
H	-0.00085	-0.00108	-0.00140	-0.00123
I	-0.00435	-0.00420	-0.00429	-0.00379
J	-0.00289	-0.00338	-0.00398	-0.00158
K	$+0.00070$	-0.00023	$+0.00103$	-0.00006
L	-0.35134	-0.37731	-0.36234	-0.38512
M	-0.02308	-0.04171	$+0.00719$	-0.05507

Table 10: PDEBench Burgers trajectory benchmark. The public trajectory uses $\nu^\star = 0.02$; training corrupts the nominal viscosity by the four mismatch ratios. Every physics row selects λ from the same four candidates by held-out validation MSE. Interior observations, validation points, and collocation points use $t \le 1$, while all reported metrics use the full $t \in [0, 2]$ trajectory.

Method	Aggregated over mismatch			Mismatch 16
	RelL2	True MSE	GDS	RelL2
data-only	0.3063 ± 0.0164	24.78 ± 3.53	1.00	0.3063
nominal PINN, tuned	0.3217 ± 0.0065	8.58 ± 0.46	1.05	0.4059
inverse PINN, tuned	0.2979 ± 0.0096	$\mathbf{6.19 \pm 0.43}$	0.97	0.3615
compatibility training, tuned	$\mathbf{0.2836 \pm 0.0165}$	11.17 ± 0.81	**0.93**	**0.2747**

observed mean RelL2 and GDS, whereas inverse PINN gives the lowest average true residual. This split shows that true-residual and prediction errors can rank the trained predictors differently. The trajectory benchmark compares the finite-budget procedures described in App. F, rather than different global predictor sets for matched joint and reduced objectives.

Table 11: Additional PDEBench Burgers trajectory results. The upper block compares all methods at the common fixed $\lambda = 0.2$ on the single trajectory. The lower block retains the four-candidate validation-selected nominal and inverse baselines for the three-trajectory setting; the common-weight comparison for that setting is Table 3.

Method	RelL2	True MSE	GDS
Single trajectory, common $\lambda = 0.2$			
data-only	0.3063 ± 0.0164	24.78 ± 3.53	1.00
nominal PINN	0.3262 ± 0.0048	8.09 ± 0.40	1.06
inverse PINN	0.2962 ± 0.0072	$\mathbf{5.24 \pm 0.35}$	0.97
compatibility training	$\mathbf{0.2775 \pm 0.0131}$	9.58 ± 0.63	**0.91**
Multiple trajectories, additional validation-selected baselines			
nominal PINN, tuned	0.1888 ± 0.0048	2.53 ± 0.12	0.828
inverse PINN, tuned	0.1799 ± 0.0049	2.64 ± 0.13	0.794

Table 12: PDEBench compressible-flow FNO with an auxiliary momentum loss, averaged over mismatch ratios $\{1, 16\}$ and, exceptionally, two data-split seeds. The physics term is the surrogate momentum diagnostic R_κ^{sur} with uncertain $\kappa := \zeta + 4\eta/3$; the surrogate MSE at $\kappa^\star$ evaluates that diagnostic at the generator coefficient and is not the exact density-weighted generator residual. All physics-informed rows use the common fixed weight $\lambda = 10^{-4}$. Boldface marks observed means, not statistical superiority.

Method	RelL2	Surrogate MSE at $\kappa^\star$	Profiled surrogate MSE	Mass MAE	GDS
data-only	0.0324 ± 0.0023	835 ± 55	809	0.0621 ± 0.0067	1.000
nominal physics	0.0366 ± 0.0020	196 ± 3	29.9	0.0564 ± 0.0045	1.129
inverse physics	$\mathbf{0.0302 \pm 0.0018}$	26.3 ± 1.2	13.7	0.0491 ± 0.0030	**0.932**
worst-case residual	0.0537 ± 0.0067	106 ± 3	87.6	0.1213 ± 0.0022	1.655
uncertainty-average	0.0432 ± 0.0013	69.0 ± 4.6	49.0	0.0979 ± 0.0001	1.333
compatibility training	0.0304 ± 0.0020	$\mathbf{21.8 \pm 2.9}$	**11.2**	$\mathbf{0.0474 \pm 0.0022}$	0.937

Table 11 retains the single-trajectory results at a common fixed weight and the additional validation-selected baselines for the multi-trajectory check. For the latter, the primary comparison in Table 3 uses $\lambda = 0.2$ for every physics method; the additional rows show the nominal and inverse outcomes when validation selection is available. Selection uses only observations in $t \leq 1$, so the validation and full-trajectory evaluation windows differ.

Compressible flow with an auxiliary momentum loss. On public compressible-flow data, compatibility training and inverse training have similar prediction errors, 0.0304 and 0.0302, at the same fixed weight. However, across the two data splits, compatibility training yields lower observed mean surrogate-residual and density-mass errors than inverse training (Table 12). The main difference is therefore in the auxiliary diagnostics.

Table 12 evaluates an uncertain auxiliary loss on a multi-variable public operator task. At the common fixed weight, compatibility training and inverse training have similar observed prediction errors, 0.0304 and 0.0302. Compatibility training has lower mean surrogate MSE at $\kappa^\star$, 21.8 against 26.3, and lower density-mass error, 0.0474 against 0.0491. Compatibility training also attains $38\times$ lower surrogate MSE than data-only while keeping prediction error below data-only. After averaging the two mismatches within each split, the RelL2 values for seeds 0 and 1 are $(0.03019, 0.03471)$ for data-only, $(0.02838, 0.03208)$ for inverse, and $(0.02838, 0.03246)$ for compatibility training. Table 13 gives the paired differences. Surrogate MSE and mass MAE are lower for compatibility training in both split averages; prediction differences are small and have opposite signs across the two splits. The residual reduction indicates closer agreement with the declared auxiliary diagnostic; RelL2 and the mass QoI assess whether the residual reduction accompanies useful predictions. The two-split check supports this descriptive comparison, not a claim of statistical superiority or equivalence to inverse training, and does not validate the exact-recovery assumptions for the generating PDE.

Table 13: Public CFD split-level results at the common $\lambda = 10^{-4}$. Each entry first averages mismatch ratios $\{1, 16\}$ within the corresponding split. Δ is compatibility training minus inverse. These split-level comparisons provide descriptive results for the two recorded splits, with no superiority or equivalence inference.

Split	Metric	Inverse	Compatibility training	Δ
0	RelL2	0.02838305	0.02837709	−0.00000596
0	Surrogate MSE at $\kappa^\star$	27.5137	18.8969	−8.6169
0	Mass MAE	0.0460999	0.0452286	−0.0008713
1	RelL2	0.03207509	0.03245705	+0.00038196
1	Surrogate MSE at $\kappa^\star$	25.0613	24.6087	−0.4526
1	Mass MAE	0.0520538	0.0495601	−0.0024937

Table 14: Non-affine reaction-law uncertainty over four mismatch levels. Every physics row uses the same four-weight validation grid. The true value is $\beta^\star = 0$ and the nominal value is $\hat{\beta} = \log \delta$. Table 16 compares global and tangent profiles at a common weight. Lower RelL2 and GDS indicate better prediction. A higher Harm value indicates a later threshold crossing as δ increases; ∞ means no harm at any tested ratio. Boldface marks the lowest observed RelL2, true MSE, and GDS means. Values that round to the same minimum are all bolded.

Method	RelL2	True MSE	GDS	Harm
data-only	0.00648 ± 0.00045	0.0604 ± 0.0080	1.00	∞
nominal PINN, tuned	0.0349 ± 0.0004	0.0268 ± 0.0007	5.39	0.25
inverse PINN, tuned	0.00656 ± 0.00009	0.00131 ± 0.00002	1.01	16
uncertainty-average, tuned	0.0963 ± 0.0008	0.0350 ± 0.0006	14.87	1
worst-case residual, tuned	0.1869 ± 0.0003	0.3556 ± 0.0150	28.85	0.25
compatibility training, tuned	$\mathbf{0.00191 \pm 0.00013}$	$\mathbf{(2.26 \pm 0.11) \times 10^{-4}}$	**0.29**	∞

G.3 Mismatch Robustness and Weak Supervision

Validation-selected non-affine comparison. Table 14 reports the validation-selected non-affine comparison summarized in Sec. 5.

Comparisons at a common physics weight. Tables 15 and 16 complement the validation-selected main comparisons. Every physics row in a table uses the same original fixed weight, allowing the affine worst-case and non-affine tangent procedures to be compared without differences in weight selection. At $\lambda = 0.2$, compatibility training improves prediction over data-only in both affine benchmarks, whereas worst-case and uncertainty-average enforcement are harmful on average. At $\lambda = 0.05$, global non-affine profiling has lower prediction error than the local tangent approximation.

The common-weight comparison in Table 15 tests whether to minimize the largest, average, or smallest residual over the uncertainty family. At $\lambda = 0.2$, reaction–diffusion GDS is 119.98 for worst-case, 117.22 for uncertainty-average, and 0.67 for compatibility training. Each benchmark's data come from one physical system, but worst-case and average losses also penalize disagreement with alternative equations in the family. These losses can penalize the true solution for violating alternative equations. Compatibility training selects the best-fitting equation, so a correct prediction receives no physics penalty when the family contains the true equation.

Mismatch sweeps. Tables 17 and 18 resolve the affine and non-affine sweeps by mismatch level. RelL2 for validation-tuned compatibility training is nearly flat across the grid, whereas every physics-informed baseline degrades most at mismatch 16.

At non-affine mismatch 16, increasing λ from 0.02 to 1.0 lowers the nominal residual from 0.260 to 0.0011 but raises RelL2 from 0.100 to 0.300 and the true residual from 0.097 to 0.320.

Sparse and noisy observations. Compatibility training also helps where the data term is weakest. We stress-test reaction–diffusion at mismatch 16 with observation counts $\{64, 128, 256, 512, 4096\}$

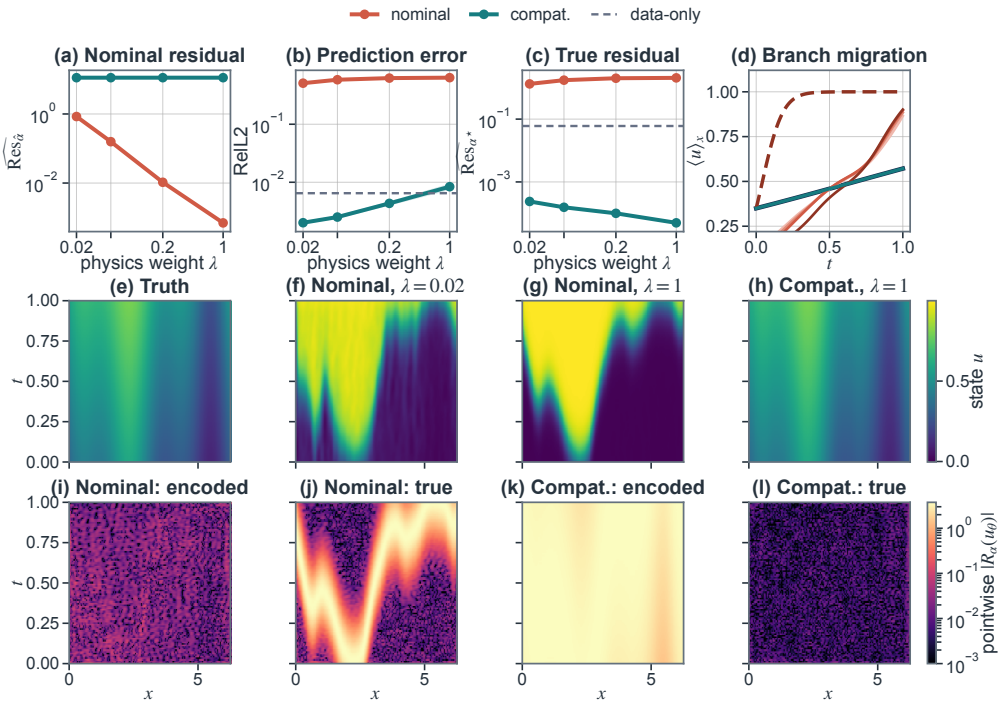


Figure 3: Anti-informative residual behavior for reaction–diffusion at mismatch 16, with $\alpha^\star = 1$ and $\hat{\alpha} = 16$. Four-seed means and one-standard-error bands cover $\lambda \in \{0.02, 0.05, 0.2, 1\}$, and dashed horizontal lines mark data-only performance. Increasing the nominal weight lowers the encoded residual from 0.843 to 7×10^{-4} while increasing RelL2 from 0.510 to 0.635 and true-residual MSE from 1.32 to 2.06. Compatibility training remains at or below RelL2 0.0084 and lowers the true residual over the same sweep. The spatial-mean trajectories use fixed seed 0: Dark blue is the truth, dashed dark red is the $\hat{\alpha} = 16$ branch, the light-to-dark red curves are nominal predictions as λ increases, and teal shows predictions from compatibility training at $\lambda = 1$. The state fields and cross-evaluations under the encoded and true equations appear on a common logarithmic scale. The cross-evaluated MSEs are 1.0×10^{-3}, 2.06, 11.3, and 7.5×10^{-5} in displayed order. Thus the nominal predictor closely satisfies the wrong equation while predictions from compatibility training stay near the true branch, matching the monotone bias floor in Theorem 1.

Table 15: Affine PINN comparisons at a common fixed $\lambda = 0.2$, aggregated over four mismatch levels and four seeds per level. All rows use the same architecture, observations, collocation budget, and 2500 predictor updates. A higher Harm value indicates a later threshold crossing as δ increases; ∞ means no harm at any tested ratio.

Method	RelL2	True MSE	GDS	Harm
Burgers				
data-only	0.0114 ± 0.0015	0.4026 ± 0.0675	1.00	∞
nominal PINN	0.0340 ± 0.0004	0.0181 ± 0.0010	2.99	4
inverse PINN	0.0181 ± 0.0007	0.0050 ± 0.0004	1.59	16
worst-case residual	0.2205 ± 0.0002	0.0762 ± 0.0007	19.38	0.25
uncertainty-average	0.1905 ± 0.0002	0.0749 ± 0.0013	16.75	0.25
compatibility training	$\mathbf{0.00700 \pm 0.00106}$	$(\mathbf{8.36 \pm 1.77}) \times \mathbf{10^{-4}}$	**0.62**	∞
Reaction–diffusion				
data-only	0.00648 ± 0.00045	0.0604 ± 0.0080	1.00	∞
nominal PINN	0.2361 ± 0.0003	0.5528 ± 0.0101	36.45	0.25
inverse PINN	0.0551 ± 0.0008	0.0219 ± 0.0007	8.50	4
worst-case residual	0.7771 ± 0.0005	0.0219 ± 0.0004	119.98	0.25
uncertainty-average	0.7592 ± 0.0007	0.0219 ± 0.0013	117.22	0.25
compatibility training	$\mathbf{0.00433 \pm 0.00039}$	$(\mathbf{9.97 \pm 0.73}) \times \mathbf{10^{-5}}$	**0.67**	∞

Table 16: Non-affine PINN comparisons at a common fixed $\lambda = 0.05$, aggregated over four mismatch levels and four seeds per level. The global profile and local tangent approximation have the same physics weight. A higher Harm value indicates a later threshold crossing as δ increases; ∞ means no harm at any tested ratio.

Method	RelL2	True MSE	GDS	Harm
data-only	0.00648 ± 0.00045	0.0604 ± 0.0080	1.00	∞
nominal PINN	0.0608 ± 0.0002	0.0441 ± 0.0006	9.39	0.25
inverse PINN	0.0112 ± 0.0001	0.00187 ± 0.00002	1.73	16
uncertainty-average	0.1429 ± 0.0004	0.0289 ± 0.0012	22.06	0.25
worst-case residual	0.2356 ± 0.0006	0.1992 ± 0.0045	36.37	0.25
tangent-projected	0.00673 ± 0.00014	$(6.5 \pm 0.1) \times 10^{-4}$	1.04	16
compatibility training	$\mathbf{0.00243 \pm 0.00018}$	$(\mathbf{1.408 \pm 0.073}) \times \mathbf{10^{-4}}$	**0.37**	∞

Table 17: Affine benchmark RelL2 by mismatch. Physics rows in the first two blocks use four-candidate validation selection; the last block retains fixed-weight nominal references. Data-only is constant in δ and serves as the reference for both protocols.

Method	Mismatch ratio δ			
	0.25	1	4	16
Burgers				
data-only		0.0114		
nominal PINN, tuned	0.0034	**0.0022**	0.0095	0.0426
inverse PINN, tuned	0.0030	0.0023	0.0050	0.0209
uncertainty-average, tuned	0.0045	0.0258	0.0890	0.2641
compatibility training, tuned	**0.0027**	0.0027	**0.0027**	**0.0022**
Reaction–diffusion				
data-only		0.0065		
nominal PINN, tuned	0.0174	**0.0019**	0.0789	0.5100
inverse PINN, tuned	**0.0020**	**0.0019**	0.0121	0.0772
uncertainty-average, tuned	0.0282	0.3210	0.9650	0.9981
compatibility training, tuned	**0.0020**	0.0020	**0.0020**	**0.0020**
Fixed-weight nominal references, $\lambda = 0.2$				
Burgers	0.0091	0.0062	0.0252	0.0956
Reaction–diffusion	0.0660	0.0044	0.2499	0.6239

Table 18: Non-affine benchmark RelL2 by mismatch, with separate validation-selected and common-weight blocks. Data-only carries no physics term, so the data-only error is constant in δ and serves as the reference for both blocks.

Method	Mismatch ratio δ			
	0.25	1	4	16
data-only	0.0065			
Validation-selected physics rows				
nominal PINN, tuned	0.0117	**0.0019**	0.0262	0.0999
inverse PINN, tuned	**0.0019**	**0.0019**	0.0028	0.0196
uncertainty-average, tuned	0.0065	0.0170	0.0977	0.2641
worst-case residual, tuned	0.0111	0.0999	0.2408	0.3957
compatibility training, tuned	**0.0019**	**0.0019**	**0.0019**	**0.0019**
Common fixed $\lambda = 0.05$, with the data-only reference above				
nominal PINN	0.0221	**0.0024**	0.0485	0.1704
inverse PINN	0.0025	**0.0024**	0.0053	0.0346
tangent-projected	0.0032	**0.0024**	0.0040	0.0172
compatibility training	**0.0024**	**0.0024**	**0.0024**	**0.0024**

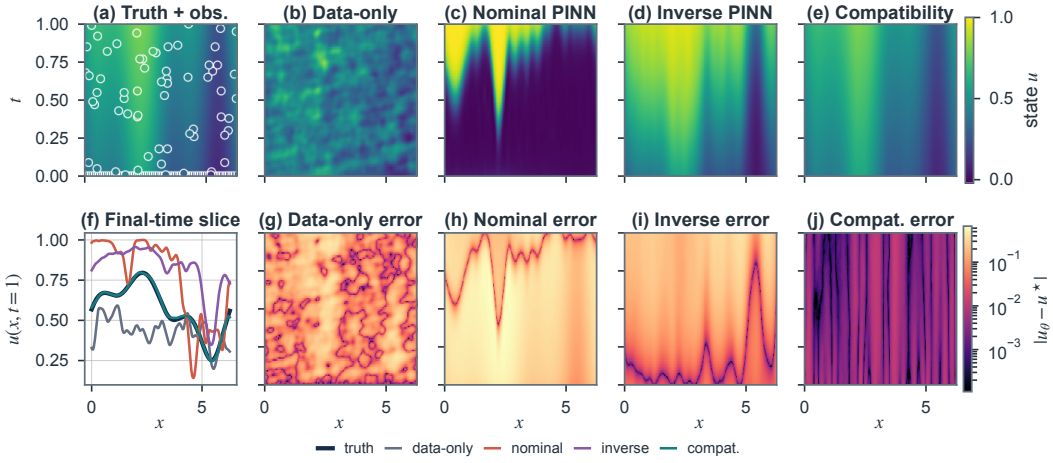


Figure 4: Sparse, noisy reaction–diffusion at mismatch 16, trained on 64 interior observations plus the full initial condition with 10% observation noise. The figure shows the ground truth, observations, and predictions from data-only, nominal, inverse, and compatibility training, followed by final-time slices and absolute-error maps on a shared logarithmic scale. For seed 0, RelL2 is 0.304, 0.771, 0.338, and 0.0114 for the four methods, respectively. All physics methods use $\lambda = 0.2$; the nominal and inverse reaction rates and the reaction rate profiled by compatibility training are approximately 15.6, 2.95, and 1.01. Table 20 reports four-seed means.

and noise levels $\{0, 1, 5, 10\}\%$. Table 19 shows the 1% noise slice, while Table 20 shows the 10% stress slice. Data-only improves with more observations but is poor in the sparse regime; the jointly trained inverse baseline also remains substantially worse than compatibility training, and nominal PINN is consistently harmful. These results show the effectiveness of the implemented profiling procedure under the stated training budget. Compatibility training keeps seed-averaged RelL2 below 0.01 for 64 to 256 observations at noise levels up to 1% and below 0.016 even at 10% noise.

For the sparsest, 10%-noise setting in Table 20, Fig. 4 shows where each method's prediction for seed 0 deviates from the true solution and which solution branch the prediction follows.

The remaining noise slices have the same ordering as Tables 19 and 20. For compatibility training, RelL2 ranges from 0.0099 down to 0.0043 at 0% noise, from 0.0114 down to 0.0045 at 5% noise, and from 0.0153 down to 0.0052 at 10% noise, while data-only ranges from 0.319 down to 0.0065, from 0.317 down to 0.0138, and from 0.316 down to 0.0263 at those same noise levels, respectively.

Table 19: Sparse and noisy reaction–diffusion at mismatch 16 and 1% observation noise. Entries are four-seed mean RelL2; all physics methods use the common fixed $\lambda = 0.2$.

n_{obs}	data-only	nominal PINN	inverse PINN	compatibility training
64	0.3189	0.8208	0.3247	**0.0099**
128	0.2302	0.6784	0.2478	**0.0095**
256	0.1528	0.6472	0.1952	**0.0071**
512	0.0992	0.6426	0.1624	**0.0058**
4096	0.0069	0.6239	0.1577	**0.0043**

Table 20: Sparse and noisy reaction–diffusion at mismatch 16 and 10% observation noise. Entries are four-seed mean RelL2; all physics methods use the common fixed $\lambda = 0.2$.

n_{obs}	data-only	nominal PINN	inverse PINN	compatibility training
64	0.3156	0.8198	0.3255	**0.0153**
128	0.2319	0.6771	0.2453	**0.0138**
256	0.1552	0.6472	0.1907	**0.0111**
512	0.1117	0.6451	0.1643	**0.0095**
4096	0.0263	0.6239	0.1543	**0.0052**

At the run level, applying the 2% harm criterion to the matched data-only run, compatibility training has zero failures across all 80 condition–seed cells, formed by 20 observation–noise settings and four seeds per setting. Inverse PINN fails in 72 cells and nominal PINN in all 80. These counts describe coverage of the tested grid, not 80 independent failure-probability trials. These results are consistent with the retained-information principle: Compatibility training can absorb represented mismatch while preserving constraints that supplement weak observations. Corollary 5 gives a local sufficient condition for this mechanism without requiring the data alone to identify the predictor. The experiment measures prediction under weak supervision rather than directly measuring the retained Hessian.

G.4 Time Extrapolation from Early Observations

Overview. With observations restricted to $t \leq 0.5$ and full-domain collocation, compatibility training achieves future-time RelL2 0.0119, compared with 0.3845 for data-only (Table 21). Removing future-domain collocation raises the error of compatibility training to 0.3100 (Table 22), with the same observation setup, collocation count, and physics weight $\lambda = 0.2$. The large extrapolation gain therefore depends on enforcing useful physics constraints in the unobserved future interval.

Sparse random observations still include the whole time interval. We therefore add a harder extrapolation test: Models observe reaction–diffusion only for $t \leq 0.5$ and are evaluated only on the unobserved future half $t \geq 0.5$. The nominal reaction rate is misspecified by a factor of 16. In Table 21, data-only extrapolates poorly and nominal physics degrades substantially at the common fixed $\lambda = 0.2$. Inverse PINN reaches RelL2 0.3809, while compatibility training reaches 0.0119, more than $30\times$ lower than both inverse PINN and data-only. The separate validation-selected baseline block retains the nominal and inverse results from the four-weight grid; the future-time errors of the selected nominal and inverse baselines remain high. The nominal residual of predictions from compatibility training remains high because the nominal equation is wrong; the true residual evaluates agreement with the generating equation.

The QoI diagnostics make the extrapolation failure more visible. On the same future-time window as Table 21, data-only has spatial-mean QoI error 0.1477 and final-time RelL2 0.4692, while nominal physics at $\lambda = 0.2$ has 0.5142 and 0.9957. Compatibility training reduces the spatial-mean QoI error and final-time RelL2 to 3.34×10^{-4} and 0.0136, respectively. Thus the gain is not confined to pointwise RelL2: The aggregate future trajectory and the final-time state are both accurately predicted.

The previous extrapolation experiment applies the PDE residual on the full space-time domain, the standard physics-informed use of a governing equation. A stricter control restricts both the

Table 21: Time extrapolation at reaction–diffusion mismatch 16. Models observe the full initial condition and 512 interior samples from $t \leq 0.5$; metrics are computed only on $t \geq 0.5$, with full-domain collocation. The primary block uses a common fixed $\lambda = 0.2$. Additional nominal and inverse baselines select from four weights using validation points in the early window, so no future-time observations enter weight selection. Boldface marks the lowest observed RelL2, true MSE, and GDS means in the primary block. Values that round to the same minimum are all bolded. Nominal MSE is unbolded because a smaller residual under the wrong equation need not mean better prediction.

Method	RelL2	True MSE	Nominal MSE	GDS
Common fixed $\lambda = 0.2$				
data-only	0.3845 ± 0.0116	4.260 ± 0.056	18.48	1.00
nominal PINN	0.9977 ± 0.0003	0.00196 ± 0.00022	0.00483	2.59
inverse PINN	0.3809 ± 0.0073	0.1383 ± 0.0022	5.955	0.99
compatibility training	$\mathbf{0.0119 \pm 0.0021}$	$(\mathbf{1.0 \pm 0.1}) \times \mathbf{10^{-4}}$	11.52	**0.03**
Additional baselines selected by validation from four candidates				
nominal PINN, tuned	0.9782 ± 0.0007	0.2405 ± 0.0402	0.699	2.54
inverse PINN, tuned	0.4824 ± 0.0039	0.3626 ± 0.0033	3.807	1.25

Table 22: The same extrapolation test with collocation also restricted to $t \leq 0.5$, which removes future-domain physics from every residual loss. The primary block uses a common fixed $\lambda = 0.2$; the additional baseline block uses the same four-candidate validation protocol as Table 21. Boldface marks the lowest observed RelL2, true MSE, and GDS means in the primary block. Values that round to the same minimum are all bolded. Nominal MSE is unbolded because a smaller residual under the wrong equation need not mean better prediction.

Method	RelL2	True MSE	Nominal MSE	GDS
Common fixed $\lambda = 0.2$				
data-only	0.3845 ± 0.0116	4.260 ± 0.056	18.48	1.00
nominal PINN	0.5615 ± 0.0080	10.38 ± 0.60	22.82	1.46
inverse PINN	0.3381 ± 0.0135	2.745 ± 0.181	18.42	0.88
compatibility training	$\mathbf{0.3100 \pm 0.0106}$	$\mathbf{1.236 \pm 0.124}$	15.79	**0.81**
Additional baselines selected by validation from four candidates				
nominal PINN, tuned	0.5218 ± 0.0236	10.40 ± 0.94	22.34	1.36
inverse PINN, tuned	0.3103 ± 0.0190	2.214 ± 0.326	17.85	0.81

supervised observations and the collocation points to $t \leq 0.5$, while evaluation remains on $t \geq 0.5$, so that no physics signal reaches the evaluation window. Table 22 measures the effect of removing that signal. At the same fixed weight, future-time RelL2 for compatibility training rises to 0.3100, compared with 0.3381 for inverse PINN and 0.3845 for data-only. The additional validation-selected inverse row reaches 0.3103. Thus the large full-domain advantage depends on residual constraints in the future window; the observed-window control gives a much smaller separation between the implemented procedures.

The aggregate diagnostics are also close between the two procedures. At $\lambda = 0.2$, compatibility training has spatial-mean QoI error 0.1032 and final-time RelL2 0.3694, compared with 0.1028 and 0.3784 for inverse PINN. Both improve final-time RelL2 over data-only. The validation-selected inverse values are 0.0952 and 0.3411, illustrating that these additional diagnostics need not rank predictors exactly as full-window RelL2 does.

G.5 Darcy Operator-Learning Validation

Scalar coefficient uncertainty. Table 23 evaluates scalar-corrupted Darcy flow when operator supervision is scarce. Compatibility training improves both RelL2 and true residual relative to data-only, while nominal, inverse, worst-case, and uncertainty-average penalties substantially worsen prediction.

Table 23: Data-limited scalar Darcy FNO with only 64 training functions, aggregated over mismatches $\{4, 16\}$. All physics methods use the common fixed $\lambda = 10^{-5}$.

Method	RelL2	True MSE	Compat. MSE	GDS
data-only	0.0903 ± 0.0058	9.337 ± 1.112	1.089	1.00
nominal physics	0.3420 ± 0.0058	0.6846 ± 0.0243	0.648	3.32
inverse physics	0.3326 ± 0.0086	$\mathbf{0.6722 \pm 0.0266}$	0.640	3.22
worst-case residual	0.5839 ± 0.0001	0.9986 ± 0.0002	0.921	8.49
uncertainty-average	0.5837 ± 0.0001	0.9976 ± 0.0003	0.903	8.48
compatibility training	$\mathbf{0.0570 \pm 0.0032}$	0.7048 ± 0.0901	**0.350**	**0.68**

Table 24: Darcy FNO with field-valued coefficient uncertainty, aggregated over three mismatch amplitudes. The model sees $a_{\mathrm{nom},i} := a_i / q_i^{\mathrm{data}}$ and the target solves with a_i. All physics methods use the common fixed $\lambda = 10^{-5}$.

Method	RelL2	True MSE	Compat. MSE	GDS
data-only	0.0197 ± 0.0005	0.6781 ± 0.0619	0.525	1.00
nominal physics	0.0201 ± 0.0005	0.0562 ± 0.0035	0.037	1.01
uncertainty-average	0.0270 ± 0.0004	0.0980 ± 0.0035	0.059	1.36
scalar compatibility training	0.0192 ± 0.0003	0.0595 ± 0.0039	0.041	0.97
field compatibility training	$\mathbf{0.0179 \pm 0.0003}$	$\mathbf{0.0416 \pm 0.0037}$	**0.035**	**0.91**

Table 25: High-resolution field-valued Darcy FNO check on a 64×64 grid with 512 training functions, over the same two mismatch amplitudes as Table 4, with common fixed $\lambda = 10^{-5}$.

Method	RelL2	True MSE	Compat. MSE	GDS
data-only	0.0190 ± 0.0002	2.8479 ± 0.0589	2.141	1.00
nominal physics	0.0218 ± 0.0018	0.0831 ± 0.0132	0.054	1.13
scalar compatibility training	0.0193 ± 0.0003	0.0728 ± 0.0024	0.047	1.01
field compatibility training	$\mathbf{0.0172 \pm 0.0002}$	$\mathbf{0.0410 \pm 0.0021}$	**0.036**	**0.90**

In data-limited scalar Darcy, compatibility training improves RelL2 from 0.0903 for data-only to 0.0570, while each other tested physics objective increases RelL2, as shown in Table 23. The stronger field-valued experiment in Table 4 tests whether the physics term needs a separate spatial correction for each input function to improve prediction when supervision is already strong.

Field-valued coefficient uncertainty. Table 24 extends scalar Darcy corruption to uncertainty that is sample-specific and spatially varying in an eight-mode basis. Scalar compatibility training improves over nominal physics only mildly, whereas field compatibility training improves both prediction and the true residual. This gap is consistent with a residual helping when the residual geometry matches the uncertainty. Under moderate misspecification, a nominal residual can remain close enough to be roughly neutral rather than harmful.

In the stronger field-valued setting of Table 4, spatial-mean pressure and boundary errors remain close to data-only: Spatial-mean pressure errors are 1.19×10^{-4} for data-only and 1.16×10^{-4} for field compatibility training, with boundary MSEs 6.0×10^{-8} and 6.2×10^{-8}, respectively.

The same ordering holds under both a finer discretization and a shift in the coefficient distribution. On the high-resolution 64×64 grid of Table 25, field compatibility training improves RelL2 from 0.0190 for data-only to 0.0172 while reducing the true residual from 2.85 to 0.0410. Under the out-of-distribution coefficient shift of Table 26, field compatibility training gives the best RelL2, 0.01366 ± 0.00030 against 0.01483 ± 0.00033 for data-only, and roughly halves the true residual, although here nominal physics attains a slightly lower mean residual diagnostic while worsening prediction.

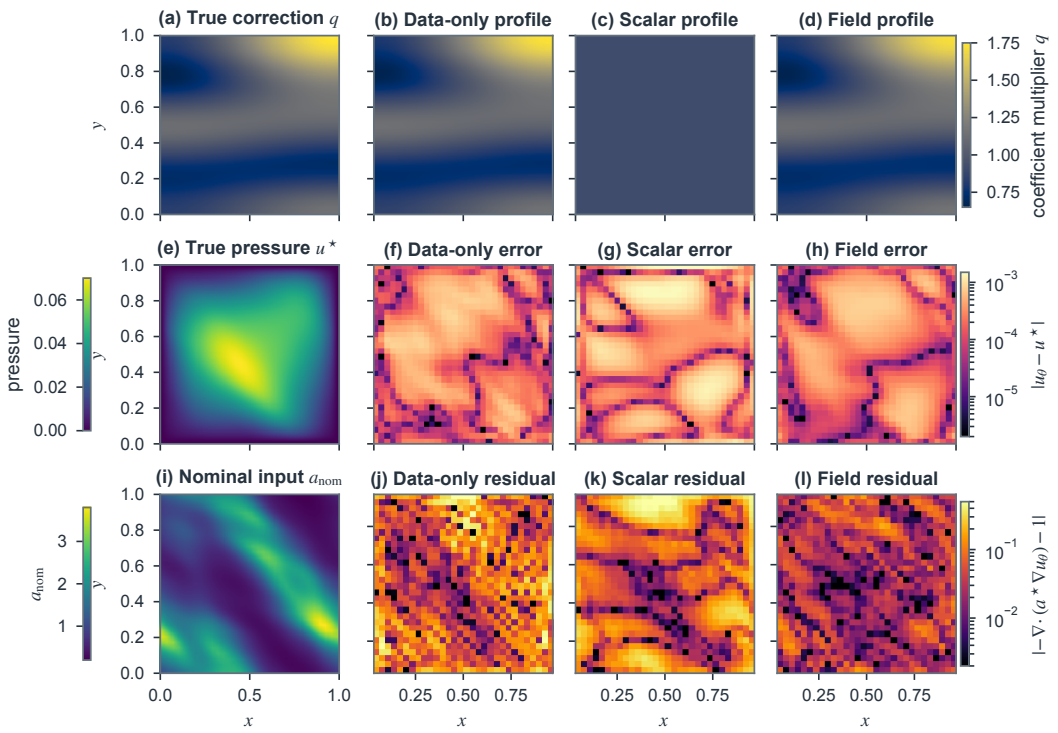


Figure 5: A representative strong field-valued Darcy FNO test function, selected by median data-only error. The FNO observes $a_{\mathrm{nom}} = a^\star/q$, where q belongs to the declared eight-mode family. The figure compares the true correction with profiles computed post hoc on data-only predictions and profiles from scalar and field compatibility training. The three profiles have root mean squared errors of 0.0117, 0.232, and 0.0105, respectively; only the field profile recovers the signed bands. The figure also shows the true pressure, absolute prediction errors, the nominal input, and true-equation residuals. The corresponding RelL2 values are 0.00788, 0.01229, and 0.00744, while the residual MSEs are 0.0237, 0.0289, and 0.00246. Error and residual maps share logarithmic scales within each row. Table 4 reports the aggregate result.

Table 26: Out-of-distribution field-valued Darcy FNO, over the same two mismatch amplitudes. The FNO is trained on smoother coefficient fields sampled with four Fourier modes and evaluated on independently generated rougher fields sampled with ten modes. All physics methods use the common fixed $\lambda = 10^{-5}$.

Method	RelL2	True MSE	Compat. MSE	GDS
data-only	0.01483 ± 0.00033	0.4554 ± 0.0669	0.377	1.00
nominal physics	0.01717 ± 0.00024	$\mathbf{0.2113 \pm 0.0100}$	**0.179**	1.15
scalar compatibility training	0.01652 ± 0.00020	0.2218 ± 0.0096	0.190	1.11
field compatibility training	$\mathbf{0.01366 \pm 0.00030}$	0.2181 ± 0.0131	0.205	**0.92**

A basis learned from separate discrepancy samples. Table 27 evaluates the construction in App. F.5. Eight PCA modes are learned from 256 independent discrepancy-prior samples and frozen before predictor training; the test multipliers are separate nonlinear transformations of prior draws. Field compatibility training lowers mean RelL2 from 0.07121 for data-only to 0.06825 and reduces the true residual by about $4.7\times$. Nominal physics has a similar observed RelL2, 0.06789, and a slightly lower true residual. The evidence here concerns the use of independently learned correction directions with an approximate family, rather than superiority to nominal training. Together with the prescribed-basis results, this check connects the mechanism under aligned uncertainty to a construction that uses no individual test corrections. The known synthetic discrepancy prior and fixed radius remain part of the experimental specification.

Table 27: Darcy FNO with an independently learned PCA discrepancy basis. The eight modes are estimated from 256 separate prior samples; the coefficient radius is fixed at $\rho = 1$. All physics methods use $\lambda = 10^{-5}$. Each entry averages mismatch amplitudes $\{0.5, 1.0\}$ within each seed, then reports the mean and standard error across four seeds.

Method	RelL2	True MSE
data-only	0.07121 ± 0.00141	1.6645 ± 0.1515
nominal physics	$\mathbf{0.06789 \pm 0.00072}$	0.3323 ± 0.0109
uncertainty-average	0.07079 ± 0.00033	$\mathbf{0.3281 \pm 0.0090}$
field compatibility training	0.06825 ± 0.00101	0.3510 ± 0.0158

Table 28: Radius sweep at reaction–diffusion mismatch 16 with fixed $\lambda = 0.2$.

Radius	Contains $\alpha^\star$	RelL2	True MSE
2	no	0.4720 ± 0.0008	0.7438
4		0.2503 ± 0.0010	0.1924
16	yes	$\mathbf{0.0043 \pm 0.0004}$	$\mathbf{1.0 \times 10^{-4}}$
64		$\mathbf{0.0043 \pm 0.0004}$	$\mathbf{1.0 \times 10^{-4}}$

G.6 Uncertainty-Set Sensitivity and Synthesis

Table 28 reports the full reaction–diffusion radius sweep summarized in Sec. 5.5.

The same containment pattern occurs beyond the affine reaction benchmark. In the fixed-$\lambda = 0.2$ Burgers radius sweep at mismatch 4, where data-only has the same RelL2 0.0114 as in Table 17, compatibility training is worse than data-only when the radius is too small, reaching RelL2 0.0154 at $\rho = 1.5$; compatibility training becomes competitive at $\rho = 2$ with RelL2 0.0108, and improves prediction once the true viscosity is included, reaching 0.0062, 0.0072, and 0.0073 at $\rho = 4$, 16, and 64. The true residual falls from 0.00263 at $\rho = 1.5$ to the minimum 0.00047 at $\rho = 4$, then remains below 0.001 at $\rho = 16$ and 64.

In the fixed-$\lambda = 0.05$ non-affine reaction radius sweep at mismatch 16, radii 2 and 4 exclude $\beta^\star = 0$ and both global compatibility training and local tangent projection remain biased; radii 16 and 64 include the truth and global compatibility training recovers RelL2 ≈ 0.0024, whereas the local tangent loss remains around 0.0172, the mismatch-16 value in Table 18, because the large nonlinear mismatch violates the local approximation.

Field-valued radius sensitivity. This sweep varies the extent of admissible coefficient corrections while keeping the eight-dimensional correction basis fixed. The radius sweep tests coverage and the effect of constraining the profile, rather than the tangent refinement of Corollary 7. The field-valued Darcy setting uses mismatch amplitude 1.0, 512 training functions, width 48, 12 modes, 80 epochs, and fixed $\lambda = 10^{-5}$. The sampled coefficient vector has Euclidean norm about $1/\sqrt{8}$, so $\rho = 0.25$ excludes that vector while $\rho \geq 0.5$ contains that vector in parameter space. Exact residual-family containment additionally requires the positivity clip in q_i^{data} to be inactive.

Table 29 shows the expected transition: There is only a marginal RelL2 gain at the too-small radius, while radii from 0.5 to 4 improve prediction and reduce the true residual by about $13\times$ relative to data-only. No collapse is observed at the larger tested radii in this ridge-stabilized, supervised setting. The lowest mean RelL2 occurs near $\rho = 1$, with small differences among the contained-radius rows. These results illustrate coverage sensitivity; the loss of identifiable curvature under added nuisance directions is quantified in Corollary 7 and isolated analytically in App. D.6.

Profile values under family enlargement. Table 30 compares post-hoc residual profiles using existing data-only Darcy records. The corruption generator uses sixteen modes throughout, while the evaluation profile uses the first $p \in \{4, 8, 16\}$ modes of the same ordered basis. The data-only training setup is fixed at 512 training functions, FNO width 48, 12 Fourier modes, batch size 32, 80 epochs, and a 32×32 grid. Each profile uses the ridge solve and radial clipping of App. F, with $\tau = 10^{-4}$ and $\rho = 1$. We average mismatch amplitudes $\{0.5, 1.0\}$ within each of four seeds.

Table 29: Field-valued Darcy FNO radius sensitivity at mismatch amplitude 1.0. Every row is field compatibility training. Data-only has RelL2 0.02284 ± 0.00033 and true residual 0.668 ± 0.029 for every radius row; the same data-only baseline is paired with each radius row so GDS is computed with matched metadata.

Radius	RelL2	True MSE	GDS
0.25	0.02248 ± 0.00092	0.0579 ± 0.0046	0.98
0.50	0.02134 ± 0.00058	0.0513 ± 0.0039	0.93
1.00	$\mathbf{0.02089 \pm 0.00061}$	0.0484 ± 0.0025	**0.91**
2.00	0.02108 ± 0.00061	$\mathbf{0.0480 \pm 0.0024}$	0.92
4.00	0.02112 ± 0.00059	$\mathbf{0.0480 \pm 0.0023}$	0.92

Table 30: Post-hoc profile values for matched data-only Darcy predictions. Only the evaluation profile changes with p; the sixteen-mode corruption and data-only training setup are fixed. Values are means and standard errors across four seed averages, each over two mismatch amplitudes. True-residual MSE is 0.7631 ± 0.0464 for every row.

Profile dimension p	RelL2	Profiled MSE
4	0.02305 ± 0.00021	0.6558 ± 0.0334
8	0.02305 ± 0.00021	0.5944 ± 0.0317
16	0.02305 ± 0.00021	0.4983 ± 0.0257

Prediction error, nominal residual, and true residual agree exactly across the three dimensions in every matched mismatch–seed cell.

The profile value falls as more correction directions become available, with unchanged prediction error. This decrease in the profile value at fixed prediction error illustrates why held-out observation error, rather than the profile value alone, should guide family selection (Sec. 4). The exact nested-family inequality in that section explains the criterion's preference for enlargement; the table reports the corresponding behavior of the stabilized approximate profiles. These post-hoc profiles are full-grid evaluation diagnostics. The post-hoc profiles measure neither a train-to-fresh-collocation gap nor the loss of identifiability from an over-flexible family.

Profile behavior. The profiled parameters are selected to evaluate the physics loss, and the parameter accuracy is reported as a diagnostic of how well the declared family captures the mismatch. Proposition 3 predicts this behavior for exact profiles under the basin and identifiability hypotheses of that proposition: The profile error is controlled by the prediction error and by the distance from the truth to the uncertainty set, so an accurate predictor pulls the profiled nuisance toward the true parameter. The ridge-stabilized field profiles reported here are therefore diagnostics analogous to, rather than direct instances of, that proposition.

The following neural-field diagnostics use the validation-selected runs of Tables 18 and 17 for both methods, with four seeds per mismatch. In the non-affine reaction benchmark, global compatibility training selects $\beta_{\mathrm{prof}} = 0$ on the evaluation grid in every selected run, matching the true $\beta^\star = 0$ and the flat RelL2 curve in Table 18. The inverse PINN's β MAE is 0.0021 at mismatch 1, 0.157 at mismatch 4, and 1.128 at mismatch 16. For reaction–diffusion, the reaction-coefficient MAE for compatibility training is 1.20×10^{-3} across the four mismatch levels, while inverse PINN's reaction-coefficient MAE grows from 9.61×10^{-4} at mismatch 1 to 2.938 at mismatch 16.

The field-valued Darcy runs show the same pattern in a higher-dimensional nuisance space. In the strong field-valued Darcy FNO setting, the MSE between the profiled eight-dimensional coefficient correction and the sampled true correction is 1.19×10^{-4} at mismatch 0.5 and 1.40×10^{-4} at mismatch 1.0 for field compatibility training, compared with 2.65×10^{-3} and 2.23×10^{-3} when the same profile diagnostic is computed on data-only predictions. Thus the RelL2 gains in Table 4 coincide with a profile that closely tracks the dominant spatial correction directions, and Fig. 5 resolves the same agreement spatially. Together with the true-residual reduction in that table, these diagnostics assess whether a lower profiled loss accompanies improved prediction and correction

accuracy. The sparse and noisy experiments separately test whether the retained constraints remain useful as observations weaken.

The field-valued radius sweep likewise improves after containment but is best near an intermediate radius (Table 29). High-resolution and coefficient-distribution-shift Darcy checks preserve the prediction ordering; the remaining benchmark tables, QoIs, and radius sweeps are collected elsewhere in App. G.

Synthesis across regimes. The reported sweeps separate three empirical regimes. Severe mismatch whose dominant nuisance direction is represented by the uncertainty family, as in reaction–diffusion, the non-affine reaction-law variant, and the sparse and noisy-data experiments, produces large RelL2 and GDS gains. In these tests, the nominal residual is anti-informative, and compatibility training avoids the resulting bias. In the strongly supervised PDEBench CFD operator with scalar uncertainty, compatibility training and joint inverse training attain comparable RelL2, while compatibility training yields a lower auxiliary residual. This empirical comparison of training procedures with a scalar nuisance does not establish that parameter identifiability distinguishes the ideal prediction objectives of compatibility training and joint inverse training. In observed-window extrapolation, no future-domain physics signal is available; compatibility training improves on data-only but the performance gap between compatibility training and inverse training is much smaller, showing that the benefit of compatibility training depends on the physics signal available over the prediction domain.

Taken together, these regimes show that the strongest gains occur when three conditions hold: The nominal residual is badly misspecified, the uncertainty set represents the dominant nuisance direction, and supervision alone does not fully determine the solution. Outcomes are closer to neutral when one of these conditions fails, as in the strongly supervised scalar operator check or when future-domain physics information is unavailable. These conditions define the operating range of the objective and indicate when profiling an uncertain auxiliary residual changes the answer and when a nominal or inverse residual already suffices.

H LIMITATIONS

Our experiments evaluate compatibility training on selected synthetic PINN benchmarks, public PDEBench tasks, and Darcy operator-learning benchmarks. These settings cover several forms of model uncertainty, but the evaluation is not exhaustive across datasets and model architectures; broader evaluation across additional tasks and models remains future work.