Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physical Fidelity of Learned Nonlinear Vehicle Dynamics: Validation Selection, Observational Bias, and Architecture-Conditioned Inductive Bias

Josafat Leal FIlho

josafat@lisha.ufsc.br

Universidade Federal de Santa Catarina

Antônio Augusto Fröhlich

Universidade Federal de Santa Catarina

Research Article

Keywords: nonlinear vehicle dynamics, physics-informed machine learning, active learning, Jacobian fidelity, finite-time dynamics, surrogate modeling

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11443334/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Physical Fidelity of Learned Nonlinear Vehicle Dynamics: Validation Selection, Observational Bias, and Architecture-Conditioned Inductive Bias

Josafat Ribeiro Leal Filho[1*] and Antônio Augusto Fröhlich[1†]

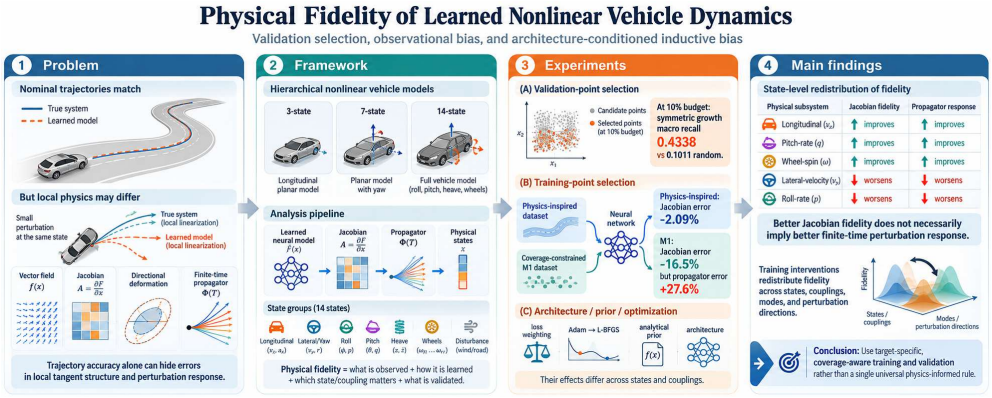
*Corresponding author(s). E-mail(s): josafat@lisha.ufsc.br;
Contributing authors: guto@lisha.ufsc.br;
†These authors contributed equally to this work.

Abstract

Learned vehicle-dynamics models can reproduce nominal trajectories while misrepresenting local tangent and finite-time behavior. We evaluate this discrepancy using analytical 3-, 7-, and 14-state nonlinear vehicle models with vector-field, Jacobian, directional-deformation, and frozen-local propagator metrics. Reference-based validation selectors substantially enrich critical-region coverage: at a 10% budget, symmetric growth reaches macro recall 0.4338 versus 0.1011 for random sampling. In fresh-seed training, physics-inspired selection reduces 14-state Jacobian error by 2.09%, below the predefined 5% practical threshold, while the 5.80% propagator reduction is not significant after Holm correction. A coverage-constrained Lagrangian selector improves Jacobian error by 16.5% but worsens propagator error by 27.6%, showing that better local derivative fidelity does not guarantee better finite-time response. Adaptive loss balancing and Adam-to-L-BFGS refinement provide no consistent gains, and architecture effects depend on learner parameterization. A post-hoc state-resolved analysis of 716 frozen checkpoints further shows that training interventions redistribute fidelity across physical states: M1 improves the longitudinal, pitch-rate, and wheel-spin equations while degrading lateral-velocity and roll-rate dynamics. These results support target-specific, coverage-aware training and validation rather than assuming that more physics-selected data or stronger physical priors improve all fidelity measures.

Keywords: nonlinear vehicle dynamics, physics-informed machine learning, active learning, Jacobian fidelity, finite-time dynamics, surrogate modeling

Graphical Abstract



1 Introduction

Learned dynamical models are increasingly used as computationally efficient surrogates for simulation, estimation, control, inverse problems, and data-driven discovery. Physics-informed learning has substantially expanded this role by combining neural approximators with differential equations, physical constraints, or structured priors [1, 2]. In parallel, nonlinear system identification and sparse discovery have emphasized that reproducing observed data is not identical to recovering the governing dynamical structure [8, 9].

For vehicle dynamics, evaluation is still often dominated by trajectory or state-prediction error. Such metrics are necessary but not sufficient for the physical-fidelity question addressed here. Two vector fields can generate similar nominal trajectories over a finite interval while having different local Jacobians, different responses to perturbations, and different finite-time amplification. This distinction becomes particularly important when a learned model is used beyond interpolation, for example in sensitivity analysis, stability assessment, controller evaluation, or safety-oriented simulation.

For a controlled nonlinear system

$$\dot{\boldsymbol{x}} = \boldsymbol{F}(\boldsymbol{x}, \boldsymbol{u}), \tag{1}$$

$\boldsymbol{x} \in \mathbb{R}^n$ denotes the vehicle state, n is the state dimension, $\boldsymbol{u} \in \mathbb{R}^7$ is the input vector, $\dot{\boldsymbol{x}} = d\boldsymbol{x}/dt$ is the state derivative with respect to physical time t, and $\boldsymbol{F} : \mathbb{R}^n \times \mathbb{R}^7 \to \mathbb{R}^n$ is the analytical reference vector field. The local tangent dynamics are characterized by

$$\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{u}) = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{x}}, \tag{2}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is the reference state Jacobian and $A_{ij} = \partial F_i / \partial x_j$ quantifies the instantaneous sensitivity of the i-th state derivative to the j-th state. A learned vector field is written as

$$\dot{\boldsymbol{x}} = \widehat{\boldsymbol{F}}_\theta(\boldsymbol{x}, \boldsymbol{u}), \tag{3}$$

where the hat denotes a learned quantity and θ collects the trainable neural-network parameters. Its state Jacobian is

$$\widehat{\boldsymbol{A}} = \frac{\partial \widehat{\boldsymbol{F}}_\theta}{\partial \boldsymbol{x}}, \tag{4}$$

which can differ from $\boldsymbol{A}$ even when the nominal vector-field prediction is accurate. Such a mismatch is relevant whenever the learned model is used in sensitivity analysis, stability assessment, closed-loop simulation, or any application in which perturbation response matters.

This work studies a second question in addition to how fidelity should be measured: *where and how should physical prior knowledge enter the learning process?* Physical knowledge can alter the distribution of observed training points, which we call *observational bias*, or it can alter the learner itself through architecture, parameterization, or an explicit physical prior, which we call *inductive bias*. These mechanisms need not have the same effect. A point that is highly informative for diagnosing a trained model may not be the point that best improves parameter estimation, and a sampling strategy that helps one architecture may harm another.

The paper therefore reports three related experiments together with a controlled extension of Experiment 2. Experiment 1 is a retrospective validation-point reduction study. Experiment 2 is a fresh-seed held-out replication that tests physics-inspired training-point selection with a fixed tanh MLP with a fitted affine output skip and vector-field loss. Its v6 extension asks whether the fixed physics/coverage allocation can be improved by explicitly constraining coverage, adapting the allocation sequentially, reweighting the physics and coverage losses, or changing the optimizer. Experiment 3 is a development-stage screening study that separates observational from inductive bias across multiple architectures. This sequence allows us to distinguish four questions:

1. Can reference-system metrics locate learned-model failure regions?
2. Does selecting training points by those physical metrics improve learned physical fidelity?
3. Can coverage-constrained or adaptive optimization improve on a frozen physics/coverage split?
4. Are the observed effects robust to, or explained by, the inductive bias of the neural learner?

Every training intervention is interpreted at two levels. At the global level we ask what happens to nominal, trajectory, Jacobian, finite-time, and stability fidelity. At the state level we ask which physical state equations, state-to-state couplings, local modes, and perturbation directions improve, worsen, or remain unchanged. The second level shows that interventions with similar global effects can act on entirely different parts of the vehicle physics.

The main contribution is therefore not a claim that one physical metric or sampling rule is universally optimal. Instead, the paper establishes a multi-scale fidelity framework and uses it to show that physical information is mechanism-dependent: validation selection, training selection, and neural inductive bias can produce different outcomes even when they are derived from the same reference dynamics.

2 Related Work

This section positions the present study at the intersection of physics-informed scientific machine learning, nonlinear system identification, active training-point selection, and architecture-induced inductive bias. The focus is not to catalogue all physics-informed methods, but to identify the parts of the literature that directly motivate the three experimental questions in this paper: whether physical structure should change how a learned dynamical model is validated, where it should be trained, and how the learner itself should be parameterized.

2.1 Physics-informed scientific machine learning

Physics-informed neural networks established a widely used paradigm in which differential-equation residuals and physical constraints are incorporated directly into neural-network training [1]. The broader physics-informed machine-learning literature subsequently generalized this idea to soft regularization, structured architectures, operator learning, and combinations of data with partially known models [2]. These highly cited studies demonstrate the value of introducing physical knowledge into learning, but they also make clear that "physics-informed" is not a single intervention: physical knowledge can enter through observations, loss functions, constraints, or architecture.

The optimization literature on PINNs is particularly relevant to the present quality-gate analysis. Wang, Teng, and Perdikaris showed that physics-informed losses can exhibit severe gradient-flow imbalances and optimization pathologies [3]. Wang, Yu, and Perdikaris later analyzed PINN training through the neural tangent kernel and showed that different loss components can converge at markedly different rates [4]. These results support a key methodological choice in this paper: a low training loss cannot by itself certify correct local physical structure. The model must be checked directly using held-out vector-field, Jacobian, and finite-time metrics.

2.2 System identification and dynamical structure

The distinction between fitting observations and identifying dynamics has a long history in nonlinear system identification. Schoukens and Ljung emphasize that nonlinear identification depends critically on excitation, model structure, data quality, and the operating region represented in the experiment [9]. In a complementary data-driven setting, Brunton, Proctor, and Kutz showed that discovering nonlinear governing equations requires an appropriate representation and sufficiently informative measurements, rather than unconstrained curve fitting [8].

These works motivate the multi-scale validation used here. Trajectory agreement measures nominal prediction, but a dynamical model intended for analysis or control should also preserve local sensitivities and perturbation response. The present study therefore evaluates not only the learned vector field but also the learned Jacobian, directional deformation, and frozen-local finite-time propagation.

2.3 Informative sampling, representativeness, and coverage

Active learning provides the closest machine-learning analogue to the observational-bias experiment. For regression, Wu identifies *informativeness*, *representativeness*, and *diversity* as distinct criteria for selecting useful samples [7]. This distinction is central to our physics-plus-coverage design: selecting only dynamically extreme points can increase physical informativeness while simultaneously reducing support of the admissible operating domain.

Within physics-informed learning, Nabian, Gladstone, and Meidani showed that nonuniform importance sampling of collocation points can materially change PINN training efficiency relative to uniform sampling [6]. Their result supports the general proposition that *where* the learner is trained matters. Our experiment tests a different, reference-dynamics-driven version of the same principle: training points are selected using Jacobian, directional, and finite-time properties rather than the current learner loss.

The present results refine this literature by showing that informative sampling is not uniformly beneficial. The sign and magnitude of the effect change with model dimension, manoeuvre, and architecture. This is consistent with the system-identification view that excitation and coverage are joint properties of an experimental design rather than interchangeable objectives [9].

2.4 Architecture and inductive bias

The learner architecture provides a second route by which physical behavior can change even when the training samples are fixed. Jagtap, Kawaguchi, and Karniadakis showed that the choice and adaptation of activation functions can substantially alter convergence and accuracy in deep and physics-informed networks [5]. More generally, residual-network approximation properties depend on depth and on the activation function or its derivatives [10]. These results provide a theoretical basis for treating activation, width, depth, and output-skip parameterization as sources of generic inductive bias rather than as neutral implementation details.

This literature also helps interpret the architecture screening reported in Experiment 3. Since the learned Jacobian is composed of products of weight matrices and activation derivatives, architecture directly constrains the class of local tangent maps that can be represented. The large descriptive variation in held-out Jacobian fidelity observed across the screened architectures is therefore qualitatively consistent with prior work showing that activation and optimization structure can materially change scientific neural-network behavior [3–5].

2.5 Gap addressed by this work

Existing studies usually examine one of the preceding mechanisms in isolation: physics-constrained learning, active sampling, nonlinear identification, or architecture design. The gap addressed here is the correspondence among them. We explicitly distinguish *validation selection*, *observational bias*, and *inductive bias*, and evaluate them using the same physical-fidelity hierarchy. This permits us to ask whether a reference metric

that is useful for locating learned-model failures is also useful for training, and whether any training effect survives changes in neural architecture.

3 Problem Formulation and Physical-Fidelity Metrics

This section defines the reference and learned dynamical systems and the multi-scale error quantities used consistently across all three experiments. The objective is to keep nominal prediction, local tangent fidelity, and frozen-local finite-time fidelity separate so that improvement in one level is not automatically interpreted as improvement in another.

3.1 Reference and learned dynamics

The study uses analytical vehicle models with $n \in \{3, 7, 14\}$ first-order states. The reduced state vectors are

$$\boldsymbol{x}_3 = \begin{bmatrix} v_x & v_y & r \end{bmatrix}^\top, \qquad \begin{aligned} \boldsymbol{x}_7 = \mathrm{col}(&v_x, v_y, r, \omega_{\mathrm{FL}}, \\ &\omega_{\mathrm{FR}}, \omega_{\mathrm{RL}}, \omega_{\mathrm{RR}}), \end{aligned} \tag{5}$$

while the 14-state model uses

$$\begin{aligned} \boldsymbol{x}_{14} = \mathrm{col}(&v_x, v_y, r, \phi, \dot{\phi}, \theta_p, \dot{\theta}_p, \\ &z, \dot{z}, \omega_{\mathrm{FL}}, \omega_{\mathrm{FR}}, \\ &\omega_{\mathrm{RL}}, \omega_{\mathrm{RR}}, F_{\mathrm{wind}}). \end{aligned} \tag{6}$$

Here v_x and v_y are longitudinal and lateral velocity, r is yaw rate, ϕ and θ_p are roll and pitch angles, $\dot{\phi}$ and $\dot{\theta}_p$ are their rates, z and $\dot{z}$ are heave displacement and heave rate, $\omega_{\mathrm{FL}}, \omega_{\mathrm{FR}}, \omega_{\mathrm{RL}}, \omega_{\mathrm{RR}}$ are the four wheel angular velocities, and F_{wind} is the wind-related state used by the 14-state implementation. The implementation stores roll rate and pitch rate internally as p and q; the notation $\dot{\phi}$ and $\dot{\theta}_p$ is used here to avoid collision with the quantile level p and the operating-point vector $\boldsymbol{q}$.

For each model, the admissible operating domain contains only physically valid state/input combinations and manoeuvre-generated states. State and input coordinates are centered and scaled using the development training pool:

$$\bar{\boldsymbol{x}} = \boldsymbol{D}_x^{-1}(\boldsymbol{x} - \boldsymbol{c}_x), \qquad \bar{\boldsymbol{u}} = \boldsymbol{D}_u^{-1}(\boldsymbol{u} - \boldsymbol{c}_u), \tag{7}$$

where $\boldsymbol{c}_x$ and $\boldsymbol{c}_u$ are coordinate-wise centers, $\boldsymbol{D}_x = \mathrm{diag}(\boldsymbol{s}_x)$ and $\boldsymbol{D}_u = \mathrm{diag}(\boldsymbol{s}_u)$ are diagonal scaling matrices, and the vectors $\boldsymbol{s}_x$ and $\boldsymbol{s}_u$ contain half-ranges of the corresponding coordinates over the development training pool. A bar therefore denotes a quantity expressed in this normalized coordinate system. With physical time left unscaled,

$$\bar{\boldsymbol{F}} = \boldsymbol{D}_x^{-1}\boldsymbol{F}, \qquad \bar{\boldsymbol{A}} = \frac{\partial \bar{\boldsymbol{F}}}{\partial \bar{\boldsymbol{x}}} = \boldsymbol{D}_x^{-1}\boldsymbol{A}\boldsymbol{D}_x. \tag{8}$$

Thus $\bar{\boldsymbol{F}}$ is the normalized vector field and $\bar{\boldsymbol{A}}$ is the normalized Jacobian. All local structural comparisons use this same basis.

3.2 Nominal fidelity

Let $\boldsymbol{q}$ denote an operating point, i.e., the complete set of coordinates used to identify one admissible state/input condition in the evaluation pool; $\boldsymbol{q}$ contains the associated $\boldsymbol{x}$ and $\boldsymbol{u}$ and any experiment-specific exogenous coordinate used by the selector. The pointwise relative normalized vector-field error is

$$E_F^{\mathrm{rel}}(\boldsymbol{q}) = \frac{\left\| \widehat{\bar{\boldsymbol{F}}}(\boldsymbol{q}) - \bar{\boldsymbol{F}}_{\mathrm{ref}}(\boldsymbol{q}) \right\|_2}{\left\| \bar{\boldsymbol{F}}_{\mathrm{ref}}(\boldsymbol{q}) \right\|_2 + \varepsilon_F}. \tag{9}$$

Here $\widehat{\bar{\boldsymbol{F}}}$ and $\bar{\boldsymbol{F}}_{\mathrm{ref}}$ are the learned and reference normalized vector fields, $\|\cdot\|_2$ is the Euclidean norm, and $\varepsilon_F > 0$ is a fixed numerical safeguard used only when the reference vector-field norm is close to zero.

Trajectory fidelity is evaluated separately. For a rollout with N_T samples at times t_k, the normalized state residual is

$$\boldsymbol{e}_x(t_k) = \boldsymbol{D}_x^{-1} \left[\widehat{\boldsymbol{x}}(t_k) - \boldsymbol{x}_{\mathrm{ref}}(t_k) \right], \tag{10}$$

where $\widehat{\boldsymbol{x}}(t_k)$ and $\boldsymbol{x}_{\mathrm{ref}}(t_k)$ are the learned and reference trajectories. The aggregate trajectory error over horizon T is

$$E_{\mathrm{traj}}(T) = \sqrt{\frac{1}{n N_T} \sum_{k=1}^{N_T} \|\boldsymbol{e}_x(t_k)\|_2^2}. \tag{11}$$

Thus N_T is the number of rollout samples and n is the state dimension. State-specific RMSE values are also retained so that normalization does not hide errors in individual states. This separation is important because a model can have small trajectory error while having poor local or finite-time fidelity.

3.3 Local Jacobian fidelity

The relative Jacobian discrepancy is

$$E_{A,\mathrm{rel}}(\boldsymbol{q}) = \frac{\left\| \widehat{\bar{\boldsymbol{A}}}(\boldsymbol{q}) - \bar{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{q}) \right\|_F}{\left\| \bar{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{q}) \right\|_F + \varepsilon_A}. \tag{12}$$

Here $\widehat{\bar{\boldsymbol{A}}}$ and $\bar{\boldsymbol{A}}_{\mathrm{ref}}$ are the learned and reference Jacobians in the normalized state basis, $\|\cdot\|_F$ is the Frobenius norm, and $\varepsilon_A > 0$ is a fixed numerical safeguard.

For a unit perturbation direction $\boldsymbol{v} \in \mathbb{R}^n$, with $\|\boldsymbol{v}\|_2 = 1$, the instantaneous longitudinal growth rate is

$$\gamma(\boldsymbol{A}, \boldsymbol{v}) = \boldsymbol{v}^\top \mathrm{sym}(\boldsymbol{A}) \boldsymbol{v}, \qquad \mathrm{sym}(\boldsymbol{A}) = \frac{\boldsymbol{A} + \boldsymbol{A}^\top}{2}. \tag{13}$$

Thus $\mathrm{sym}(\boldsymbol{A})$ is the symmetric part of the Jacobian and γ measures instantaneous stretching or contraction along $\boldsymbol{v}$. The corresponding transverse reorientation quantity is

$$g_F^C(\boldsymbol{A}, \boldsymbol{v}) = 4 \left\| (\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{A}\boldsymbol{v} \right\|_2^2, \tag{14}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix and $(\boldsymbol{I} - \boldsymbol{v}\boldsymbol{v}^\top)$ projects the tangent action $\boldsymbol{A}\boldsymbol{v}$ onto the subspace orthogonal to $\boldsymbol{v}$.

The directional analysis uses the frozen probe set

$$\mathcal{V}_{\min} = \{\boldsymbol{e}_i\}_{i=1}^n \cup \left\{ \frac{\boldsymbol{e}_i + \boldsymbol{e}_j}{\sqrt{2}} : 1 \le i < j \le n \right\},$$
$$K = |\mathcal{V}_{\min}| = \frac{n(n+1)}{2}. \tag{15}$$

where $\boldsymbol{e}_i$ is the i-th canonical basis vector and K is the number of probe directions; $K = 105$ for $n = 14$. Let $\boldsymbol{\gamma}(\boldsymbol{A}) \in \mathbb{R}^K$ and $\boldsymbol{g}_F(\boldsymbol{A}) \in \mathbb{R}^K$ collect γ and g_F^C over those directions. The relative directional discrepancies are

$$D_\gamma^{\mathrm{rel}}(\boldsymbol{q}) = \frac{\|\boldsymbol{\gamma}(\widehat{\boldsymbol{A}}) - \boldsymbol{\gamma}(\bar{\boldsymbol{A}}_{\mathrm{ref}})\|_2}{\|\boldsymbol{\gamma}(\bar{\boldsymbol{A}}_{\mathrm{ref}})\|_2 + \varepsilon_\gamma}, \tag{16}$$

$$D_{g_F}^{\mathrm{rel}}(\boldsymbol{q}) = \frac{\|\boldsymbol{g}_F(\widehat{\boldsymbol{A}}) - \boldsymbol{g}_F(\bar{\boldsymbol{A}}_{\mathrm{ref}})\|_2}{\|\boldsymbol{g}_F(\bar{\boldsymbol{A}}_{\mathrm{ref}})\|_2 + \varepsilon_{g_F}}, \tag{17}$$

where ε_γ and ε_{g_F} are fixed numerical safeguards. The Experiment 3 development evaluator additionally stores RMS directional differences; whenever those development-only quantities are used, their aggregation convention is stated explicitly.

3.4 Frozen-local finite-time fidelity

The finite-time metrics used in the reported experiments are *frozen-local*, not trajectory-varying tangent propagators. At operating point $\boldsymbol{q}$, the local state-transition matrix over horizon T is

$$\boldsymbol{\Phi}(\boldsymbol{q}; T) = \exp\big(\bar{\boldsymbol{A}}(\boldsymbol{q})T\big), \qquad T = 0.5 \text{ s}, \tag{18}$$

where $\exp(\cdot)$ is the matrix exponential, $\boldsymbol{\Phi} \in \mathbb{R}^{n \times n}$ maps an infinitesimal normalized state perturbation from the beginning to the end of the frozen-local horizon, and T is fixed to 0.5 s in the reported experiments. The propagator discrepancy is the absolute Frobenius error

$$E_\Phi(\boldsymbol{q}) = \left\| \widehat{\boldsymbol{\Phi}}(\boldsymbol{q}; T) - \boldsymbol{\Phi}_{\mathrm{ref}}(\boldsymbol{q}; T) \right\|_F, \tag{19}$$

where $\widehat{\boldsymbol{\Phi}}$ and $\boldsymbol{\Phi}_{\mathrm{ref}}$ are generated from the learned and reference normalized Jacobians, respectively. The frozen-local finite-time Lyapunov quantity is

$$\lambda_{\mathrm{FT}}(\boldsymbol{q}; T) = \frac{1}{T} \log \sigma_{\max}(\boldsymbol{\Phi}(\boldsymbol{q}; T)), \tag{20}$$

where $\sigma_{\max}(\cdot)$ denotes the largest singular value. Its absolute learned/reference discrepancy is

$$E_{\mathrm{FTLE}}(\boldsymbol{q}) = \left|\widehat{\lambda}_{\mathrm{FT}}(\boldsymbol{q};T) - \lambda_{\mathrm{FT,ref}}(\boldsymbol{q};T)\right|. \tag{21}$$

3.5 State-resolved fidelity metrics

The scalar metrics above answer whether the learned model is globally faithful. To ask *which part of the vehicle physics* a training intervention changes, the same normalized quantities are resolved by physical state. The ordering of $\boldsymbol{x}_{14}$ in Eq. (6) was verified against the implemented state vector (`vx`, `vy`, `yaw_rate`, `roll`, `roll_rate`, `pitch`, `pitch_rate`, `heave`, `heave_rate`, the four wheel speeds, and `crosswind_force`). The fixed physical groups are $G_{\mathrm{LONG}} = \{v_x\}$, $G_{\mathrm{LAT/YAW}} = \{v_y, r\}$, $G_{\mathrm{ROLL}} = \{\phi, \dot{\phi}\}$, $G_{\mathrm{PITCH}} = \{\theta_p, \dot{\theta}_p\}$, $G_{\mathrm{HEAVE}} = \{z, \dot{z}\}$, $G_{\mathrm{WHEELS}} = \{\omega_{FL}, \omega_{FR}, \omega_{RL}, \omega_{RR}\}$, and $G_{\mathrm{DIST}} = \{F_{\mathrm{wind}}\}$. Jacobian-block tables keep the finer partition of the earlier post-hoc analysis, in which lateral velocity and yaw rate are separate groups, so that previously reported block values remain directly comparable.

For state i the nominal vector-field error is

$$E_{F,i} = \mathrm{mean}_{\boldsymbol{q}} \left|\widehat{\bar{F}}_i(\boldsymbol{q}) - \bar{F}_{\mathrm{ref},i}(\boldsymbol{q})\right|, \tag{22}$$

which, because Eq. (7) scales every state by its half-range, is the per-state contribution to the gate error E_F^{gate}. Row i of $\bar{\boldsymbol{A}}$ contains the sensitivities of $\dot{x}_i$ to every state; it is the local linearization of the *differential equation of state i*. Column j contains the sensitivities of every derivative to x_j; it describes how *state j influences the rest of the vehicle*. The corresponding relative errors are

$$E_{A,\mathrm{row}}(i) = \mathrm{mean}_{\boldsymbol{q}} \frac{\|\widehat{\bar{\boldsymbol{A}}}_{i,:} - \bar{\boldsymbol{A}}_{\mathrm{ref},i,:}\|_2}{\|\bar{\boldsymbol{A}}_{\mathrm{ref},i,:}\|_2 + \varepsilon}, \tag{23}$$

$$E_{A,\mathrm{col}}(j) = \mathrm{mean}_{\boldsymbol{q}} \frac{\|\widehat{\bar{\boldsymbol{A}}}_{:,j} - \bar{\boldsymbol{A}}_{\mathrm{ref},:,j}\|_2}{\|\bar{\boldsymbol{A}}_{\mathrm{ref},:,j}\|_2 + \varepsilon}, \tag{24}$$

and group versions replace the row (column) by the rows (columns) of a group and the Euclidean by the Frobenius norm. The distinction matters physically: a poorly learned yaw equation ($E_{A,\mathrm{row}}(r)$ large) is a different defect from a poorly learned influence of yaw rate on lateral and roll motion ($E_{A,\mathrm{col}}(r)$ large).

For the canonical perturbation $\delta\boldsymbol{x}_0 = \boldsymbol{e}_j$ of one normalized unit of state j, the frozen-local response after $T = 0.5$ s is $\boldsymbol{\Phi}_{\mathrm{ref}}\boldsymbol{e}_j$, and the directional propagator error is

$$E_{\Phi,\mathrm{dir}}(j) = \frac{\|\widehat{\boldsymbol{\Phi}}\boldsymbol{e}_j - \boldsymbol{\Phi}_{\mathrm{ref}}\boldsymbol{e}_j\|_2}{\|\boldsymbol{\Phi}_{\mathrm{ref}}\boldsymbol{e}_j\|_2 + \varepsilon}. \tag{25}$$

It answers whether a perturbation of physical state j is propagated through the vehicle correctly. Group directions use the equal-weight unit vector over the states of a group. Because a single spurious unstable eigenvalue can make $\widehat{\boldsymbol{\Phi}}$ explode, the pointwise median

over operating points is the primary summary and the mean is reported alongside it. Selected state-to-state entries $|\widehat{\Phi}_{ij} - \Phi_{\mathrm{ref},ij}|$ (response of i to a perturbation of j) were declared before any state-level result was computed: $v_y \to r$, $r \to v_y$, $r \to \phi$, $\phi \to v_y$, $v_y \to \dot{\phi}$, $\omega_{FL} \to v_x$, $\omega_{RL} \to v_x$, $\omega_{FL} \to \phi$, $v_x \to \theta_p$, $v_x \to \omega_{FL}$, and $F_{\mathrm{wind}} \to v_y$.

Finally, the learned Jacobian of Eq. (36) is split exactly into its prior, affine, and neural parts, $\widehat{\boldsymbol{A}} = \boldsymbol{A}_{\mathrm{prior}} + \boldsymbol{A}_{\mathrm{aff}} + \boldsymbol{A}_{\mathrm{NN}}$ (the sum reproduces the protocol Jacobian to 4×10^{-16}). Two row-wise quantities describe where the model is actually nonlinear: the neural share $r_{\mathrm{NN},i} = \mathrm{mean}_{\boldsymbol{q}} \, \|\boldsymbol{A}_{\mathrm{NN}}[i,:]\| / (\|\widehat{\boldsymbol{A}}[i,:]\| + \varepsilon)$, and the state dependence of the learned row,

$$V_i = \left[\frac{\mathrm{mean}_{\boldsymbol{q}} \, \|\widehat{\boldsymbol{A}}_{i,:}(\boldsymbol{q}) - \overline{\widehat{\boldsymbol{A}}}_{i,:}\|_2^2}{\mathrm{mean}_{\boldsymbol{q}} \, \|\widehat{\boldsymbol{A}}_{i,:}(\boldsymbol{q})\|_2^2} \right]^{1/2}, \tag{26}$$

where the overbar is the mean over operating points. $V_i = 0$ means that the learned equation of state i is affine over the domain; comparing V_i with the reference value V_i^{ref} shows whether the learner represents the state dependence that the reference dynamics actually have. At the seven Jacobian entries supplied by $\boldsymbol{F}_{\mathrm{prior}}$, the missing-prior fraction $m_{ij} = -\langle \widehat{\bar{A}}_{ij} - \bar{A}_{\mathrm{ref},ij}, P_{ij} \rangle / \langle P_{ij}, P_{ij} \rangle$, with P the scaled prior Jacobian and the inner product taken over operating points, measures which share of the exact prior term the learned model lacks ($m = 1$) or reproduces ($m = 0$); under $I1$ a positive m means that the learned residual cancels that share of the supplied prior.

Five reference rows—ϕ, θ_p, z, $\dot{z}$, and F_{wind}—are constant over the operating domain ($V^{\mathrm{ref}} = 0$): their equations are linear in the implemented model. Every affine-skip learner reproduces them to at most 1.3% relative error, so percentage changes between two such negligible errors carry no physical meaning. They are reported as "exactly represented" rather than as relative changes.

3.6 Critical regions

For validation-point selection, a critical region is defined from the *learned-model discrepancy*, not from the reference selector. For error mechanism k,

$$\mathcal{C}_k(p) = \{ \boldsymbol{q} : E_k(\boldsymbol{q}) \geq Q_p(E_k) \}, \tag{27}$$

where k indexes the learned-model error mechanism, $E_k(\boldsymbol{q})$ is its pointwise error, $p \in (0,1)$ is the chosen tail quantile, and $Q_p(E_k)$ is the empirical p-quantile of E_k over the evaluation set. Thus $\mathcal{C}_k(p)$ contains points in the high-error tail of mechanism k. If $\mathcal{S}_m(B)$ is the set selected by reference-only metric m under an integer point budget B, the corresponding recall is

$$R_{m \to k}(B) = \frac{|\mathcal{S}_m(B) \cap \mathcal{C}_k|}{|\mathcal{C}_k|}. \tag{28}$$

In Eq. (28), $|\cdot|$ denotes set cardinality and the intersection counts selected points that are also critical for mechanism k. Macro recall is the arithmetic mean of the mechanism-specific recalls, whereas "Recall any" uses the union of the mechanism-specific critical sets.

4 Training Metrics, Validation Metrics, and Sources of Bias

This section establishes the conceptual bridge between the physical quantities used to prioritize operating points and the discrepancies used to validate the trained model. It also separates observational bias, which changes the training distribution, from inductive bias, which changes the learner or its objective.

4.1 Reference-only informativeness

The reference-only training and validation selectors include quantities such as Jacobian variation I_J, symmetric growth, directional γ, transverse g_F^C, non-normality, and finite-time amplification. For example, let

$$\bar{\boldsymbol{q}} = \boldsymbol{D}_q^{-1}(\boldsymbol{q} - \boldsymbol{c}_q) \tag{29}$$

denote normalized operating coordinates, where $\boldsymbol{c}_q$ is the coordinate-wise center and $\boldsymbol{D}_q$ is the corresponding diagonal scale matrix. Jacobian variation is then

$$I_J(\boldsymbol{q}) = \left\| \nabla_{\bar{\boldsymbol{q}}} \bar{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{q}) \right\|_F , \tag{30}$$

where $\nabla_{\bar{\boldsymbol{q}}}$ denotes differentiation with respect to the normalized operating coordinates. Therefore I_J measures how rapidly the reference local tangent dynamics vary across the operating domain; it is not a learned-model error or a direct measure of danger.

4.2 Three levels of metric correspondence

A useful distinction is:

$$\boxed{\begin{aligned} \rho_{mk}, \mathrm{AP}_{mk} &: \quad \text{Does reference metric } m \text{ predict error } k? \\ R_{m\to k} &: \quad \text{Does it locate critical error efficiently?} \\ \Delta_{m\to k} &: \quad \text{Does training there reduce held-out error?} \end{aligned}} \tag{31}$$

Here ρ_{mk} is the Spearman rank correlation between reference score s_m and learned error E_k, AP_{mk} is the average precision obtained by ranking points with s_m against membership in the critical set for error k, $R_{m\to k}$ is defined by Eq. (28), and $\Delta_{m\to k}$ denotes the training-effect contrast on held-out error k after selecting training points using metric m. Pointwise association and validation recall can be strong even when the training effect is weak or reversed, because training changes the fitted model and therefore changes the error field.

4.3 Observational and inductive bias

We use *observational bias* to denote a change in the training-point distribution while the learner and loss remain fixed. Experiment 2 isolates this mechanism by changing the selected operating points under a fixed MLP learner and vector-field objective. In

Experiment 3, the observational comparison is random selection ($O0$) versus a frozen physics-plus-coverage policy ($O4$).

The inductive comparison in Experiment 3 is not a physics loss and not a hard constraint. It is an additive analytical prior on the model output. The normalized coordinates $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{u}}$ are defined in Eq. (7). Let $\boldsymbol{\sigma}_F \in \mathbb{R}^n$ contain the per-component standard deviation of $\boldsymbol{F}$ over the development training pool, let $\boldsymbol{W}_{\text{aff}}$ denote the frozen affine output map fitted by ridge regression on the selected training points, let $\boldsymbol{N}_\theta(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) \in \mathbb{R}^n$ denote the nonlinear neural residual, and let $\odot$ denote componentwise multiplication. The no-prior learner is

$$\widehat{\boldsymbol{F}}_{I0}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{\sigma}_F \odot \left[\boldsymbol{W}_{\text{aff}} \begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{u}} \\ 1 \end{bmatrix} + \boldsymbol{N}_\theta(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) \right]. \tag{32}$$

The inductive condition adds the fixed prior

$$\widehat{\boldsymbol{F}}_{I1}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{F}_{\text{prior}}(\boldsymbol{x}) + \boldsymbol{\sigma}_F \odot \left[\boldsymbol{W}_{\text{aff}} \begin{bmatrix} \bar{\boldsymbol{x}} \\ \bar{\boldsymbol{u}} \\ 1 \end{bmatrix} + \boldsymbol{N}_\theta(\bar{\boldsymbol{x}}, \bar{\boldsymbol{u}}) \right], \tag{33}$$

and trains on the same scaled vector-field MSE after subtracting $\boldsymbol{F}_{\text{prior}}$ from the target. Thus $I1$ changes the parameterization of the function to be learned, not the objective weights. The prior contains only exact analytical terms already present in the reference equations:

$$F_{\text{prior},v_x} = r v_y, \qquad F_{\text{prior},v_y} = -r v_x, \tag{34}$$

for all model levels, and, for the 14-state model,

$$F_{\text{prior},\phi} = \dot{\phi}, \qquad F_{\text{prior},\theta_p} = \dot{\theta}_p, \qquad F_{\text{prior},z} = \dot{z}. \tag{35}$$

In Eqs. (34)–(35), $F_{\text{prior},j}$ denotes the component of the analytical prior associated with state derivative j; r is yaw rate, v_x and v_y are longitudinal and lateral velocity, and $\dot{\phi}$, $\dot{\theta}_p$, and $\dot{z}$ are roll-rate, pitch-rate, and heave-rate states. All remaining prior components are zero. The prior uses no labels, tyre model, or fitted physical parameters, and the learned residual can in principle cancel it. It is therefore a soft parameterization bias rather than a hard physical constraint.

The corresponding learned Jacobian contains three contributions,

$$\widehat{\boldsymbol{A}} = \frac{\partial \boldsymbol{F}_{\text{prior}}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{F}_{\text{aff}}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{F}_{\text{NN}}}{\partial \boldsymbol{x}}, \tag{36}$$

where $\boldsymbol{F}_{\text{aff}}$ denotes the vector-field contribution of the frozen affine output map and $\boldsymbol{F}_{\text{NN}}$ denotes the contribution of the nonlinear network $\boldsymbol{N}_\theta$; the prior derivative is absent under $I0$. The affine term is fitted once by ridge regression on the selected training points and then frozen. The neural term depends on activation derivatives, depth, and width, so generic neural architecture remains a second source of inductive bias even when the physical prior is fixed.

The loss weights are identical under $I0$ and $I1$: only vector-field MSE is optimized; Jacobian, directional-growth, g_F, and propagator losses are zero. Consequently, Experiment 3 tests a clean output-prior intervention and does not test Jacobian- or propagator-supervised physics-informed learning.

4.4 Informative, representative, and invalid training points

The notion of a "good training point" requires care because three different properties are involved. A physically informative point is one at which the reference dynamics exhibit a mechanism of interest strongly, for example rapid Jacobian variation, large symmetric growth, pronounced directional reorientation, non-normality, finite-time amplification, or proximity to tyre saturation. A representative point serves a different purpose: even if its physics score is moderate, it improves support of the admissible operating domain and reduces the amount of extrapolation required between the regions covered by the training set. An invalid point is qualitatively different from both. It lies outside the declared physical domain or carries a target that is inconsistent with the reference system. Consequently, a low physics score does not imply a poor training sample, and a high physics score does not imply that the resulting training set is sufficient.

Experiment 3 makes this distinction concrete through the frozen $O4$ selection policy, where $O4$ denotes the deterministic physics-plus-coverage observational treatment and B denotes the total number of training points allowed by the current budget. With $\alpha = 0.25$, the fraction assigned to the physics component, the remaining $1 - \alpha = 0.75$ is assigned to deterministic coverage:

$$\mathcal{S}_{O4} = \mathrm{unique}\left(\mathcal{S}_{\mathrm{phy}}(0.25B) \cup \mathcal{S}_{\mathrm{cov}}(0.75B)\right), \tag{37}$$

where $\mathcal{S}_{\mathrm{phy}}$ is the physics-selected subset, $\mathcal{S}_{\mathrm{cov}}$ is the deterministic-coverage subset, and $\mathrm{unique}(\cdot)$ removes duplicate point identities. Deterministic refill is applied whenever overlap reduces the union below B. The two subsets are generated independently and only then merged. The physics component ranks candidates through the empirical-percentile composite

$$s_{\mathrm{comp}}(\boldsymbol{q}) = \frac{1}{6}\sum_{m=1}^{6} P_m(s_m(\boldsymbol{q})), \tag{38}$$

where $s_m(\boldsymbol{q})$ is the raw reference-only score of metric m, $P_m(\cdot)$ maps that score to its empirical percentile within the development training pool, and $m = 1, \ldots, 6$ indexes I_J, symmetric growth, g_F, frozen-local finite-time amplification, non-normality, and the tyre-transition score. High-scoring points are selected greedily while maintaining, whenever possible, a minimum Euclidean separation $\rho_{\mathrm{sel}} = 0.10$ in the normalized operating coordinates $\bar{\boldsymbol{q}}$. The subscript "sel" distinguishes this diversity threshold from the Spearman coefficient ρ_{mk} in Eq. (31). The coverage component, in contrast, uses no physics score; it distributes the budget across manoeuvres and selects points near equally spaced progression anchors. In this way the two parts address different deficiencies of a finite training set: the physics component seeks dynamically informative regions, whereas the coverage component seeks broad support of the admissible domain.

The development sweep used to freeze this policy is especially important for its interpretation. The tested values $\alpha \in \{0.25, 0.50, 0.75, 1.00\}$, where α denotes the fraction of the budget assigned to the physics component, produced mean relative changes of -3.2%, -11.0%, -19.4%, and -20.7%, respectively, against random sampling on the frozen multi-metric development criterion. Thus, every tested mixture was worse than random, and $\alpha = 0.25$ was retained because it was the least harmful candidate, not because it was an optimum. This result is central to the interpretation of the later training experiments: increasing the fraction of high-score physics points does not produce a monotonic improvement in the learner and can instead reduce the balance between local informativeness and global domain support.

The comparison across architectures and inductive-bias conditions remains valid because the selected point identities are held fixed within each comparison. Architecture and the $I0/I1$ condition are deliberately excluded from the selection cache key, and the implementation refuses to compare fits whose selection hashes differ. Therefore, changes attributed to inductive bias or architecture are not hidden changes in the training data.

Finally, genuinely problematic data must be distinguished from merely low-score data. Let $\mathcal{Q}_{\mathrm{adm}}$ denote the physically admissible operating-point set and let $\boldsymbol{y}(\boldsymbol{q})$ denote the supervised target attached to point $\boldsymbol{q}$, which is the reference vector-field target in the training experiments. The excluded set can be written conceptually as

$$\mathcal{Q}_{\mathrm{bad}} = \left\{ \boldsymbol{q} : \begin{array}{l} \boldsymbol{q} \notin \mathcal{Q}_{\mathrm{adm}} \text{ or} \\ \boldsymbol{y}(\boldsymbol{q}) \text{ is inconsistent with} \\ \text{the reference model} \end{array} \right\}. \tag{39}$$

Thus $\mathcal{Q}_{\mathrm{bad}}$ denotes invalid or target-inconsistent points, not merely valid points with a low physics score. Physically impossible states, invalid state/input combinations, corrupted targets, or mislabeled derivatives change the learning problem itself rather than merely allocating budget to a less informative region. These cases are excluded by the admissibility rules and are not part of the robustness claims of this work. The empirical evidence therefore supports a training-data principle in which correctness, physical informativeness, and operating-domain coverage are complementary requirements, while their relative weighting must be determined for the target metric, learner architecture, and available data budget rather than assumed a priori.

5 Experimental Design

This section describes four experiments and their different evidential roles, together with the v6 extension of Experiment 2. Experiment 1 is retrospective and diagnostic. Experiment 2 provides the main fresh-seed held-out training comparison under a fixed learner; its v6 extension evaluates adaptive physics/coverage allocation, loss balancing, and optimizer choice without changing the underlying vector-field learning task. Experiment 3 is a development-stage screening study of architecture and inductive-bias interactions. Experiment 4 defines a physics-informed optimal-design-of- experiments protocol in which admissible initial conditions and control excitations are optimized to generate informative reachable trajectories and physics collocation points.

The experiments form a deliberate hierarchy. Experiment 1 asks where a trained model should be validated. Experiment 2 asks whether physics-informed observations improve training. The v6 extension asks whether the manner in which physical informativeness and coverage are combined can improve the fixed-allocation result. Experiment 3 asks whether those effects depend on the inductive bias of the learner. Experiment 4 asks a different question: whether the data-generation process itself can be optimized so that the operational data domain and a complementary physics-collocation domain jointly produce a more structurally faithful neural dynamics model under the same acquisition budget.

5.1 Experiment 1: validation-point selection

Experiment 1 is a retrospective validation-point study designed to determine whether reference-system quantities can concentrate the operating points at which an already trained model exhibits large physical-fidelity errors. The comparison includes random selection, `fixed_point`, trajectory-density sampling, I_J, the development-best single selector `symmetric_growth`, and the frozen combined selector C_{top3}. Ten independent neural seeds are evaluated, while the random baseline is averaged over 100 random subsets per budget so that each selector is compared at the same achieved sample count. Because the held-out outcomes had already been inspected before the regenerated completion run, these results are treated as retrospective evidence and are used to motivate the later training experiments rather than as a new untouched confirmatory test.

The notation C_{top3} refers specifically to the legacy development-selected combined validation selector from Experiment 1. The archived completion material records that the underlying composite quantities were converted to empirical percentiles and that the combined selector was frozen using development manoeuvres only. The artifacts available for the present manuscript, however, do not preserve the identities of the three constituent metrics or the exact algebraic form by which they were combined. For that reason, C_{top3} is reported exactly as a historical frozen selector and is neither reconstructed nor retuned. This point is important because C_{top3} is not the later six-metric s_{comp} used by `physics_composite` and `physics_plus_coverage`. Experiment 1 therefore evaluates a legacy combined *validation* selector, whereas Experiments 2 and 3 evaluate explicitly defined *training* selectors.

5.2 Experiment 2: observational-bias training study

Experiment 2 is the v5 physics-inspired training study. The run reports 2,400 fits with 10 fresh neural seeds and fresh held-out operating points, using the 3-, 7-, and 14-state models across five data budgets. It compares three physics-agnostic strategy labels and twelve physics-inspired labels. Two physics labels, `finite_time` and `FTLE`, are monotone transformations that select identical point sets, so the twelve labels represent eleven distinct physics selections. The frozen family-level analysis is retained exactly as run, which means that this duplicated selection enters the reported physics-family median twice. The learner is a fixed two-hidden- layer tanh MLP with a fitted affine output skip, trained with vector-field loss only. Within each seed, the median result

over the physics-inspired strategies is compared with the mean result over the agnostic strategies. The primary observational-bias comparison is the 14-state model at the 5% training budget. The design deliberately keeps architecture and loss fixed, following the active-learning principle that a sampling effect should be isolated from learner changes [7].

The held-out operating points and seeds are fresh, but the manoeuvre families had been inspected in earlier work. We therefore describe Experiment 2 as a fresh-seed held-out replication rather than as an untouched benchmark.

5.3 Experiment 2b: adaptive physics–coverage allocation and optimizer axis

The v6 extension keeps the frozen v5 learner and held-out protocol but changes how the finite training budget is allocated and optimized. Its primary cell is the 14-state model at a 5% budget with Adam and ten independent network seeds. Let $\mathcal{P}$ denote the development training pool and let

$$B = \lceil b|\mathcal{P}| \rceil \tag{40}$$

be the number of selected samples for fractional budget b. For candidate point i, $\boldsymbol{q}_i$ denotes normalized operating coordinates and $s_i \in [0,1]$ is the stable rank-normalization of the development physics-composite score. All fits retain the two-hidden-layer $[64, 64]$ tanh residual MLP, 500 epochs, batch size 128, learning rate 10^{-3}, and vector-field-only supervision used by v5.

The frozen v5 comparator allocates approximately $\alpha_{\text{base}}B$ samples to diversity-constrained physics selection and the remaining samples to deterministic coverage, with $\alpha_{\text{base}} = 0.25$. This quantity is the configured allocation parameter, not the realized fraction of PHY-labelled samples in the final set. For any selected set S, define the realized PHY fraction as

$$f_{\text{PHY}}(S) = \frac{|S_{\text{PHY}}|}{|S|}, \tag{41}$$

where S_{PHY} denotes the points labelled as originating from the physics-selection contribution. Because the physics and coverage subsets are merged, duplicates are removed, and any shortfall is refilled, in general $f_{\text{PHY}}(S) \neq \alpha_{\text{base}}$. In the primary 14-state, 5% cell, the frozen baseline has $\alpha_{\text{base}} = 0.25$ but $f_{\text{PHY,base}} = 0.262$. Deterministic coverage is the physics-agnostic component: it allocates the available budget across manoeuvres and trajectory progression in a reproducible manner, without using a reference physics score or learned-model error. The v6 extension replaces this fixed fraction by an explicit coverage functional. For a fixed set $\mathcal{A}$ of 2000 evenly spaced pool anchors,

$$F(S) = \sum_{a \in \mathcal{A}} \max_{i \in S} \exp\left(-\frac{\|\boldsymbol{q}_a - \boldsymbol{q}_i\|_2^2}{2h^2} \right), \tag{42}$$

$$h = \text{median}_{a \in \mathcal{A}}\, \text{NNdist}(a), \tag{43}$$

$$\widehat{C}(S) = \frac{F(S)}{F(S_{\mathrm{cov}})}, \tag{44}$$

where S is a selected set, S_{cov} is the deterministic-coverage set of the same cardinality, h is the median nearest-neighbour distance among the anchors, and $\widehat{C}(S)$ is coverage normalized to that baseline. The facility-location functional is monotone and submodular, so it rewards sets that place at least one selected point close to each anchor while avoiding redundant concentration.

The first adaptive selector, denoted M1 (`physics_coverage_lagrangian`), follows the constrained-learning view in which performance objectives are optimized subject to explicit requirements rather than absorbed into a single unconstrained loss [14]. In the present application, the objective is average physics informativeness and the explicit requirement is minimum operating-domain coverage:

$$\max_{|S|=B} \frac{1}{B} \sum_{i \in S} s_i \quad \text{subject to} \quad \widehat{C}(S) \geq \rho_C, \qquad \rho_C = 0.90. \tag{45}$$

The implemented scalarization is

$$J_\alpha(S) = \alpha \frac{1}{B} \sum_{i \in S} s_i + (1-\alpha)\widehat{C}(S), \qquad \lambda = \frac{1-\alpha}{\alpha}, \tag{46}$$

and lazy greedy is used under the cardinality constraint. Because the facility-location term is monotone submodular and the physics-score term is modular, this choice is consistent with the classical greedy approximation framework for monotone submodular maximization under a cardinality constraint [15]. The selected $\alpha^\star$ is the largest value whose greedy set still satisfies Eq. (45), obtained by bisection on $[0, 1]$ with tolerance 10^{-4}. The constrained-learning reference motivates the Lagrangian formulation, whereas the bisection used here is the operational procedure adopted in this experiment to identify the largest feasible trade-off weight. Thus M1 does not prescribe a fixed fraction of physics points; it permits strong physics concentration only when the formal coverage constraint remains satisfied.

5.3.1 Operational step-by-step procedure for coverage-constrained physics selection

The procedure combines three established ideas rather than introducing a new optimization theory: explicit constrained learning through Lagrangian trade-offs [14], greedy optimization of a monotone submodular coverage objective [15], and a one-dimensional feasibility search for the largest trade-off weight that still satisfies the coverage constraint. The M1 selector can be reproduced as the following sequence. The purpose of the procedure is to maximize the physical informativeness of the selected training set without allowing the set to lose a prescribed level of geometric coverage. The physics score and the coverage functional are therefore computed independently and are combined only during subset optimization.

Step 1. Construct the admissible development pool and normalize its operating coordinates. Start from the development training pool $\mathcal{P} = \{\boldsymbol{q}_i\}_{i=1}^{N_{\mathcal{P}}}$, after all physical admissibility filters have been applied. For each candidate, transform the operating coordinates to the common normalized basis,

$$\bar{\boldsymbol{q}}_i = \boldsymbol{D}_q^{-1}(\boldsymbol{q}_i - \boldsymbol{c}_q), \tag{47}$$

where $\boldsymbol{c}_q$ and $\boldsymbol{D}_q$ are fixed from development data. The fractional training budget b is converted to an integer number of points,

$$B = \lceil b|\mathcal{P}| \rceil . \tag{48}$$

The normalization must be frozen before selection because all subsequent distance and coverage calculations depend on this geometry.

Step 2. Compute one reference-only physics score for every candidate. For each $\bar{\boldsymbol{q}}_i$, evaluate the reference-system metrics used by the physics-composite selector and combine their empirical-percentile values into the composite score $s_{\mathrm{comp}}(\boldsymbol{q}_i)$. The resulting scores are converted to the stable rank-normalized quantity

$$s_i \in [0, 1], \tag{49}$$

where larger s_i means that candidate i is more informative according to the frozen physics criterion. No learned-model error is used at this stage.

Step 3. Build the coverage reference independently of the physics score. Construct the fixed anchor set $\mathcal{A} = \{\bar{\boldsymbol{q}}_a\}_{a=1}^{N_{\mathcal{A}}}$, with $N_{\mathcal{A}} = 2000$ in the reported experiment. Determine the kernel length scale

$$h = \mathrm{median}_{a\in\mathcal{A}}\, \mathrm{NNdist}(a), \tag{50}$$

and generate the deterministic-coverage set S_{cov} with the same budget B. For any candidate subset S, evaluate

$$F(S) = \sum_{a\in\mathcal{A}} \max_{i\in S} \exp\left(-\frac{\|\bar{\boldsymbol{q}}_a - \bar{\boldsymbol{q}}_i\|_2^2}{2h^2}\right), \tag{51}$$

followed by the normalized coverage

$$\widehat{C}(S) = \frac{F(S)}{F(S_{\mathrm{cov}})}. \tag{52}$$

Thus $\widehat{C}(S) = 1$ denotes the facility-location coverage achieved by the deterministic-coverage reference at the same cardinality. The reported constraint requires $\widehat{C}(S) \geq \rho_C = 0.90$.

Step 4. For a trial value of α, greedily construct the highest-value set. For fixed $\alpha \in [0,1]$, initialize $S = \varnothing$ and maximize

$$J_\alpha(S) = \alpha \frac{1}{B} \sum_{i \in S} s_i + (1-\alpha)\widehat{C}(S). \tag{53}$$

At each greedy iteration, compute the marginal value of an unselected candidate i,

$$\Delta J_\alpha(i \mid S) = \frac{\alpha}{B} s_i + (1-\alpha)\left[\widehat{C}(S \cup \{i\}) - \widehat{C}(S)\right], \tag{54}$$

and select

$$i^\star = \arg\max_{i \notin S} \Delta J_\alpha(i \mid S). \tag{55}$$

Then update $S \leftarrow S \cup \{i^\star\}$ and repeat until $|S| = B$. Equation (54) makes the role of the selection explicit: a point is valuable if it has a large physics score, if it covers a region that is not yet represented, or if it contributes to both. The implementation uses lazy greedy to avoid recomputing every marginal gain from scratch after every accepted point.

Step 5. Check the explicit coverage constraint and adapt α. After constructing S_α, evaluate $\widehat{C}(S_\alpha)$. If

$$\widehat{C}(S_\alpha) \geq 0.90, \tag{56}$$

the trial value is feasible and the search can move toward a larger α, thereby increasing the emphasis on physics informativeness. If $\widehat{C}(S_\alpha) < 0.90$, the trial is infeasible and the search moves toward a smaller α. Starting from the interval $[0,1]$, the implementation applies bisection until the interval width is below 10^{-4}. This feasibility search is an implementation-specific way of tuning the Lagrangian trade-off parameter within the constrained-learning formulation [14]; the subset generated at each trial value is obtained with the monotone-submodular greedy construction discussed above [15]. The final value is

$$\alpha^\star = \max\left\{\alpha : \widehat{C}(S_\alpha) \geq 0.90\right\}, \tag{57}$$

and the selected training set is

$$S^\star = S_{\alpha^\star}. \tag{58}$$

The important point is that $\alpha^\star$ is an objective weight, not a prescribed fraction of physics-labelled samples.

Step 6. Freeze the selected set before neural-network training and evaluate it on held-out data. Once $S^\star$ is obtained, neither its point identities nor $\alpha^\star$ are changed using held-out outcomes. The neural learner is trained on $S^\star$ using the same architecture, loss, optimizer settings, and seed protocol as the comparator. After the learner-quality gate, the trained model is evaluated on fresh held-out operating points using $E_{A,\mathrm{rel}}$, E_Φ, and the other declared fidelity metrics. This last step is essential because satisfying the geometric coverage

constraint does not guarantee improvement in every dynamical quantity. In the primary 14-state, 5% result reported in Sec. 6.7, M1 reduces $E_{A,\mathrm{rel}}$ by 16.5% while increasing E_Φ by 27.6%. The selection algorithm must therefore be validated against the physical quantity required by the downstream application rather than accepted solely because it satisfies Eq. (52).

Operationally, the procedure can be summarized as

$$\begin{aligned}\text{admissible pool} &\to \text{normalized physics scores}\\ &\to \text{coverage anchors} \to \text{greedy } J_\alpha \text{ selection}\\ &\to \text{coverage test} \to \text{bisection on } \alpha\\ &\to S^\star \to \text{freeze} \to \text{train} \to \text{evaluate}.\end{aligned} \tag{59}$$

The second adaptive selector, M2 (`physics_coverage_primal_dual`), uses the primal–dual view of constrained active learning developed by Elenter, Naderializadeh, and Ribeiro [16], together with a multiplicative dual update in the spirit of Generalized ELBO with Constrained Optimization (GECO) [17]. It divides the budget into five sequential rounds. The first round uses deterministic coverage, after which a learner is fitted for 200 epochs and each subsequent batch maximizes J_{α_k} conditional on the points already selected. Let e_{cov} and e_{phy} denote the pre-update errors on the newly selected coverage- and physics-labelled points. The multiplier and physics weight are updated as

$$c_k = \log e_{\mathrm{cov}} - \log e_{\mathrm{phy}} - \log(1+\varepsilon), \tag{60}$$

$$\lambda_{k+1} = \mathrm{clip}_{[10^{-3},10^{3}]}\left(\lambda_k e^{\eta c_k}\right), \tag{61}$$

$$\alpha_k = \frac{1}{1+\lambda_k}, \tag{62}$$

with $\varepsilon = 0.10$, $\eta = 1$, and $\lambda_0 = 1$. Here the adaptive quantity is not a fixed subset fraction: λ_k is the dual multiplier and $\alpha_k = 1/(1+\lambda_k)$ is the corresponding score-versus-coverage trade-off weight. The use of dual variables to encode constraint pressure is consistent with primal–dual constrained active learning [16], while the multiplicative exponential update follows the GECO-style constraint-balancing principle [17]. A known implementation limitation is that if a batch contains no physics-labelled points, the previous c_k is reused; a positive reused constraint can drive λ_k to its upper clip. Results from M2 are therefore interpreted with this limitation explicitly in view.

A second axis holds the frozen-alpha point set fixed and adapts the relative training emphasis of its physics-labelled (PHY) and coverage-labelled (COV) subsets. The common objective is

$$\begin{aligned}\mathcal{L}_t &= w_{\mathrm{phy}} L_{\mathrm{phy}} + w_{\mathrm{cov}} L_{\mathrm{cov}},\\ L_g &= \frac{1}{n_g d}\sum_{i\in g} \|\boldsymbol{r}_i - \mathrm{MLP}(\boldsymbol{z}_i)\|_2^2 .\end{aligned} \tag{63}$$

where $g \in \{\mathrm{PHY}, \mathrm{COV}\}$, n_g is the number of samples from group g, d is the output dimension, $\boldsymbol{z}_i$ is the normalized network input, and $\boldsymbol{r}_i$ is the scaled residual target after subtraction of the frozen affine component. The weights are normalized so that $w_{\mathrm{phy}} + w_{\mathrm{cov}} = 2$. Five adaptive weighting rules are evaluated. The `dualweight` rule uses the GECO constrained-weighting principle [17]; learning-rate annealing (`lra`) follows the gradient-balancing strategy of Wang, Teng, and Perdikaris [3]; `gradnorm` follows Chen et al. [18]; `softadapt` follows Heydari, Thompson, and Mehmood [19]; and `relobralo` follows Bischof and Kraus [20]. These methods estimate or update the relative PHY/COV loss weights during training; unlike M1 and M2, they do not change the selected point identities.

Finally, the optimizer axis compares the frozen Adam loop with the same Adam training followed by full-batch L-BFGS-B refinement for 500 iterations (memory 20 and gradient tolerance 10^{-9}). This Adam-to-L-BFGS sequence is included as an optimizer-control axis motivated by recent analyses of PINN loss landscapes and hybrid first-/second-order training [21]. L-BFGS-B alone is retained as a descriptive 3-state check and is not part of the 14-state primary comparison. All v6 outcomes use fresh held-out operating points and the same scaled Jacobian and frozen-local propagator definitions as Eqs. (12) and (19). Seed-level paired differences are defined as strategy minus comparator, so negative values indicate lower error. Confidence intervals use a 2000-resample percentile bootstrap over the ten seeds, significance uses an exact two-sided Wilcoxon signed-rank test, and Holm correction is applied within each comparison family.

5.4 v5 learner-quality-gate amendment

The learner-quality gate is defined on development validation points using the implementation-specific scaled vector-field error

$$E_F^{\mathrm{gate}}(\boldsymbol{q}) = \left\| \left(\widehat{\boldsymbol{F}}(\boldsymbol{q}) - \boldsymbol{F}_{\mathrm{ref}}(\boldsymbol{q}) \right) \oslash \boldsymbol{s}_x \right\|_2 , \tag{64}$$

where $\boldsymbol{s}_x$ is the state-scale vector introduced in Eq. (7). This quantity is distinct from the relative fidelity metric E_F^{rel} in Eq. (9). A fit passes when

$$\begin{aligned} \mathrm{mean}\left(E_F^{\mathrm{gate}}\right) &\le 0.95 \min(E_{F,\mathrm{zero}}, E_{F,\mathrm{mean}}), \\ \mathrm{median}(E_{A,\mathrm{rel}}) &< 0.90. \end{aligned} \tag{65}$$

where $E_{F,\mathrm{zero}}$ is the mean gate error of the zero vector-field predictor,

$$E_{F,\mathrm{zero}} = \mathrm{mean}_{\boldsymbol{q}} \left\| \boldsymbol{F}_{\mathrm{ref}}(\boldsymbol{q}) \oslash \boldsymbol{s}_x \right\|_2 ,$$

and $E_{F,\mathrm{mean}}$ is the corresponding error of the predictor that returns the mean training target. The operators $\mathrm{mean}(\cdot)$ and $\mathrm{median}(\cdot)$ are taken over the development validation points.

The gate amendment is part of the scientific record and must be disclosed. Under the previous v4 interpretation, 21 fits associated with primary-budget physics-inspired

selections failed the quality gate: all seeds of `gamma` for the 14-state, 2% condition; all seeds of `symmetric_growth` for the 7-state, 10% condition; and one additional `symmetric_growth` fit. Every agnostic fit passed.

The v5 analysis preserved the existing thresholds and all completed fits. The agnostic fits are used to certify base-learner adequacy, while the 21 failed physics-inspired fits remain in the results as evidence of strategy-induced failure rather than being excluded. The amendment was fixed before the v5 held-out data were generated.

At the supplementary 1% budget, 216 fits failed the unchanged thresholds. The 3- and 14-state 1% cells are treated descriptively. The 7-state 1% cell failed only on physics fits and is inference-eligible under the v5 rule; this is the corrected interpretation used in the present manuscript.

5.5 Experiment 3: observational × inductive bias × architecture screening

Experiment 3 is a development-stage screening study, not a confirmatory test. It uses the 14-state model, 2% and 10% budgets, three development seeds, 500 fixed epochs, and 14 architectures. The four factorial cells are $O0I0$ (random/no prior), $O4I0$ (physics-plus-coverage/no prior), $O0I1$ (random/analytical prior), and $O4I1$ (both interventions). All 336 fits passed the quality gate, and no test data were used.

The architectures vary activation, depth, width, and presence of the fitted affine output skip. The baseline $A0$ is a 2-layer, width-64 tanh MLP with the affine skip; `plain64` removes that skip; activation variants use SiLU, GELU, softplus, or ReLU; depth variants use 4 or 6 hidden layers; width variants use 128 or 256 units; and `depth4_matched` uses width 40 so its total parameter count approximately matches $A0$. The largest tested model, `medium256`, has four width-256 hidden layers. Here "skip" means the frozen affine output model, not a stack of residual blocks. The affine map is fitted by ridge regression with $\lambda = 10^{-5}$ before neural training; `plain64` has no affine map.

Training is otherwise fixed: Adam with learning rate 10^{-3}, batch size 128, no scheduler, no early stopping, no weight decay, float64 arithmetic, Xavier-uniform initialization with the last-layer weights scaled by 0.01, and identical selected-point IDs in every comparison that is intended to isolate architecture or inductive bias. Evaluation uses the same 600 development points for every fit.

For any evaluated error metric E, let $E(Oa, Ib)$ denote the seed-level error obtained with observational treatment Oa and inductive condition Ib. Here $O0$ is random point selection, $O4$ is the frozen physics-plus-coverage selection, $I0$ is the no-prior learner, and $I1$ is the analytical-prior learner. Positive contrasts are defined to mean lower error:

$$O = E(O0, I0) - E(O4, I0), \tag{66}$$

$$I = E(O0, I0) - E(O0, I1), \tag{67}$$

$$J = E(O0, I0) - E(O4, I1), \tag{68}$$

$$O \times I = J - O - I \tag{69}$$

$$= E(O4, I0) + E(O0, I1) - E(O0, I0) - E(O4, I1). \tag{70}$$

The scalar O is the observational main-effect contrast, I is the inductive main-effect contrast, J is the joint contrast relative to the $O0I0$ baseline, and $O \times I$ is their difference-in-differences interaction. A positive interaction means that the combined treatment lowers the error more than the sum of the two separate effects; it can nevertheless occur when both main effects are harmful and the combination is merely *less harmful*, so the interaction sign must be interpreted together with the four absolute cell errors.

Because only three seeds are available, all Experiment 3 results are reported descriptively. No confirmatory architecture ranking or general interaction claim is assigned.

5.6 Experiment 4: physics-informed optimal experiment design and collocation

Protocol status. The following subsection defines the planned Experiment 4 protocol. It should be interpreted as an experimental design, not as a completed empirical result; numerical claims should be added only after the protocol has been executed on fresh seeds and held-out operating points.

The preceding experiments select training points from an already generated candidate pool. Experiment 4 moves the intervention one level earlier and optimizes *how the dynamical system is excited*. The objective is to combine two complementary domains. The *operational domain* $\mathcal{D}_{\text{op}}$ is defined by physically reachable trajectories generated from admissible initial conditions and control signals and represents the state–input combinations that the learner is expected to reproduce from data. The *physics-collocation domain* $\mathcal{C}_{\text{phys}}$ is a set of admissible, reachable state–input points selected because the reference dynamics indicate that they contain high structural information. The neural model is trained on both domains: supervised data constrain nominal behavior over $\mathcal{D}_{\text{op}}$, while physics-informed collocation constrains the learned dynamics at optimized points in $\mathcal{C}_{\text{phys}}$.

5.6.1 Reachable optimal design of experiments

Let an experiment be parameterized by

$$\boldsymbol{\xi} = (\boldsymbol{x}_0,\, \boldsymbol{c}_u,\, \boldsymbol{p})\,, \tag{71}$$

where $\boldsymbol{x}_0$ is an admissible initial state, $\boldsymbol{c}_u$ parameterizes the control trajectory, and $\boldsymbol{p}$ contains optional scenario parameters such as road friction, road curvature, load, or external disturbance conditions when they are part of the declared model. The control trajectory $\boldsymbol{u}_{\boldsymbol{\xi}}(t)$ is represented by a low-dimensional piecewise-linear or spline basis so that steering, longitudinal actuation, and their rates remain inside the same admissible bounds used by the reference model. The resulting state trajectory is constrained by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}_{\text{ref}}\left(\boldsymbol{x}(t), \boldsymbol{u}_{\boldsymbol{\xi}}(t); \boldsymbol{p}\right), \tag{72}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{73}$$

$$\boldsymbol{x}(t) \in \mathcal{X}_{\text{adm}}, \qquad \boldsymbol{u}_{\boldsymbol{\xi}}(t) \in \mathcal{U}_{\text{adm}}. \tag{74}$$

Optimizing controls rather than arbitrary state coordinates is important: every selected state must be dynamically reachable under an admissible excitation rather than merely satisfying static bounds.

For a reachable operating point $\boldsymbol{q} = (\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{p})$, let $s_{\mathrm{comp}}(\boldsymbol{q}) \in [0, 1]$ denote the same stable rank-normalized physics-composite score used in the earlier training study. It combines the declared reference quantities associated with Jacobian variation, instantaneous symmetric growth, transverse reorientation, finite-time amplification, non-normality, and tyre-transition behavior. Let $\widehat{C}(S)$ denote the coverage functional in Eq. (44). For the trajectory generated by design $\boldsymbol{\xi}$, define

$$S(\boldsymbol{\xi}) = \{\boldsymbol{q}_{\boldsymbol{\xi}}(t_j)\}_{j=1}^{N_\xi}.$$

The primary optimal-excitation problem is

$$\begin{aligned}\boldsymbol{\xi}^\star = \arg\max_{\boldsymbol{\xi}\in\Xi_{\mathrm{adm}}} \quad & \omega_s \frac{1}{N_\xi}\sum_{j=1}^{N_\xi} s_{\mathrm{comp}}\left(\boldsymbol{q}_{\boldsymbol{\xi}}(t_j)\right) \\ & + \omega_C \Delta\widehat{C}\left(S(\boldsymbol{\xi}) \mid S_{\mathrm{prev}}\right) - \omega_u R_u(\boldsymbol{\xi}) - \omega_{\Delta u} R_{\Delta u}(\boldsymbol{\xi}),\end{aligned} \tag{75}$$

subject to Eq. (74). Here $\Delta\widehat{C}(S(\boldsymbol{\xi}) \mid S_{\mathrm{prev}})$ is the incremental coverage obtained by the candidate experiment relative to the experiments already accepted, R_u penalizes unnecessary control effort, and $R_{\Delta u}$ penalizes excessive control-rate variation. The weights $\omega_s, \omega_C, \omega_u, \omega_{\Delta u}$ are fixed on development data before the final experiment. A constrained multi-objective formulation may be used instead of the scalarized form, for example by maximizing physics informativeness subject to a minimum incremental coverage requirement. The important methodological constraint is that the acquisition objective is computed only from the reference dynamics and declared operating-domain constraints; held-out neural-model errors are not used to choose the experiment.

Multiple excitation experiments are generated sequentially. At round k,

$$\boldsymbol{\xi}_k^\star = \arg\max_{\boldsymbol{\xi}\in\Xi_{\mathrm{adm}}} J_{\mathrm{DoE}}\left(\boldsymbol{\xi} \mid S_{1:k-1}\right), \tag{76}$$

after which the new reachable samples are added to $S_{1:k}$. This sequential construction prevents every optimized maneuver from collapsing onto the same high-score region and makes the control-design problem explicitly coverage aware.

5.6.2 Operational data and optimized physics collocation

The operational dataset is

$$\mathcal{D}_{\mathrm{op}} = \{(\boldsymbol{q}_i, \boldsymbol{y}_i)\}_{i=1}^{N_{\mathrm{op}}}, \qquad \boldsymbol{y}_i = \boldsymbol{F}_{\mathrm{ref}}(\boldsymbol{q}_i), \tag{77}$$

where the operating points are sampled from the accepted reachable trajectories. In a later simulator or real-vehicle study, $\boldsymbol{y}_i$ may instead be the measured or numerically estimated state derivative. Thus $\mathcal{D}_{\mathrm{op}}$ represents the empirical operational domain rather than an unconstrained Cartesian state grid.

A larger candidate collocation pool $\mathcal{Q}_{\mathrm{DoE}}$ is constructed from the optimized trajectories and admissible local perturbations that pass the same physical-domain filters used elsewhere in the manuscript. The final physics-collocation set is then selected by

$$\max_{\substack{\mathcal{C}\subset\mathcal{Q}_{\mathrm{DoE}}\\ |\mathcal{C}|=B_{\mathrm{phys}}}} \frac{1}{B_{\mathrm{phys}}}\sum_{\boldsymbol{q}\in\mathcal{C}} s_{\mathrm{comp}}(\boldsymbol{q}) \quad \text{subject to} \quad \widehat{C}(\mathcal{C}) \geq \rho_C. \tag{78}$$

This is the same separation between physical informativeness and operating- domain coverage introduced by M1, but the candidate pool is now generated by optimized excitation rather than fixed in advance. Consequently, Experiment 4 tests two coupled interventions: *where the system is excited* and *which reachable points are retained for physics-informed training.*

5.6.3 Physics-informed neural training

The learner receives nominal information from $\mathcal{D}_{\mathrm{op}}$ and structural information from $\mathcal{C}_{\mathrm{phys}}$. The supervised operational-data term is

$$\mathcal{L}_{\mathrm{data}} = \frac{1}{N_{\mathrm{op}}}\sum_{(\boldsymbol{q}_i,\boldsymbol{y}_i)\in\mathcal{D}_{\mathrm{op}}} \left\| \widehat{\boldsymbol{F}}_\theta(\boldsymbol{q}_i) - \boldsymbol{y}_i \right\|_2^2. \tag{79}$$

Because the present study has access to analytical reference dynamics, the primary physics-collocation residual can be written as

$$\mathcal{L}_{\mathrm{phys}} = \frac{1}{B_{\mathrm{phys}}}\sum_{\boldsymbol{q}_j\in\mathcal{C}_{\mathrm{phys}}} \left\| \widehat{\boldsymbol{F}}_\theta(\boldsymbol{q}_j) - \boldsymbol{F}_{\mathrm{ref}}(\boldsymbol{q}_j) \right\|_2^2. \tag{80}$$

This term should be described precisely as *reference-supervised physics collocation* in the analytical study. In a partially known or experimental system, Eq. (80) can be replaced by residuals of the known balance laws, kinematic identities, or other exact physical constraints without assuming that the full reference vector field is available.

The central structural-loss ablation directly addresses a limitation of the earlier vector-field-only studies. Define

$$\mathcal{L}_A = \frac{1}{B_{\mathrm{phys}}}\sum_{\boldsymbol{q}_j\in\mathcal{C}_{\mathrm{phys}}} \frac{\left\| \widehat{\bar{\boldsymbol{A}}}_\theta(\boldsymbol{q}_j) - \bar{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{q}_j) \right\|_F^2}{\left\| \bar{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{q}_j) \right\|_F^2 + \varepsilon_A}, \tag{81}$$

$$\mathcal{L}_\Phi = \frac{1}{B_{\mathrm{phys}}}\sum_{\boldsymbol{q}_j\in\mathcal{C}_{\mathrm{phys}}} \left\| \exp\left(\widehat{\bar{\boldsymbol{A}}}_\theta(\boldsymbol{q}_j)T\right) - \exp\left(\bar{\boldsymbol{A}}_{\mathrm{ref}}(\boldsymbol{q}_j)T\right) \right\|_F^2. \tag{82}$$

The complete candidate objective is

$$\mathcal{L}_{\mathrm{E4}} = \mathcal{L}_{\mathrm{data}} + \lambda_F\mathcal{L}_{\mathrm{phys}} + \lambda_A\mathcal{L}_A + \lambda_\Phi\mathcal{L}_\Phi. \tag{83}$$

The weights are fixed on development data and then frozen. They must not be retuned on the final held-out set. The propagator-aware term is included because the v6 M1 result shows that lower Jacobian error does not by itself guarantee lower finite-time propagator error.

5.6.4 Factorized comparison and fairness constraints

Experiment 4 is designed to separate the effect of optimized data generation from the effect of a stronger training objective. The first comparison axis holds Eq. (83) fixed and changes only the source of collocation points:

1. random reachable collocation;
2. deterministic coverage collocation;
3. M1 selection from the original fixed candidate pool; and
4. optimized-excitation DoE followed by Eq. (78).

The second axis holds the optimized DoE/collocation set fixed and changes only the loss:

1. data plus vector-field physics residual ($\lambda_A = \lambda_\Phi = 0$);
2. data plus vector-field and Jacobian losses ($\lambda_A > 0, \lambda_\Phi = 0$); and
3. data plus vector-field, Jacobian, and propagator losses ($\lambda_A > 0, \lambda_\Phi > 0$).

This factorization prevents an improvement caused by optimized excitation from being attributed to the addition of Jacobian or propagator supervision.

All comparisons use the same neural architecture, optimizer, initialization seeds, number of operational samples, number of collocation points, and training-step budget. Acquisition comparisons additionally use the same number of trajectories or the same total simulated time so that an optimized control policy cannot obtain an advantage by simply querying the reference system more often. The primary study uses the 14-state model with fresh network seeds and a fresh held-out operating-point set. A 10% acquisition budget is a natural primary condition because Experiment 1 showed its strongest validation-selector enrichment at 10%; because that observation motivated the new design, all Experiment 4 confirmatory statistics must use newly generated held-out data rather than the retrospective Experiment 1 points. The 5% budget is retained as a secondary condition to test whether the benefit persists when acquisition is more restrictive.

5.6.5 Primary hypotheses and evaluation

The primary Experiment 4 question is whether jointly optimizing excitation and physics collocation improves structural fidelity under a matched acquisition budget. The preregistered primary contrasts should compare the optimized DoE arm with deterministic coverage and with M1. For both $E_{A,\mathrm{rel}}$ and E_Φ, the same decision rule used in the v6 analysis can be retained: the paired mean difference must favor the optimized method, the Holm-adjusted p-value must be below 0.05, and the mean relative reduction

must reach the predefined 5% practical threshold. Secondary outcomes include vector-field error, rollout error, the learner-quality gate, realized coverage, control effort, and the fraction of optimized points lying in each maneuver or physical-mechanism regime.

The experiment should therefore answer three distinct questions. First, does optimal excitation discover a more useful reachable state–control domain than sampling from a fixed pool? Second, after the operational domain has been generated, does optimization-based physics collocation improve the neural model relative to random or coverage-only collocation? Third, does explicit Jacobian or propagator supervision remove the local-versus-finite-time trade-off observed for M1? These questions must be evaluated separately so that "better design of experiments", "better collocation", and "better loss" are not collapsed into a single treatment.

5.7 Statistical interpretation

The independent neural seed is the inferential unit. For Experiment 2, the primary analysis uses paired seed-level differences and the predefined multiple-testing correction. Statistical significance and practical importance are treated separately. The primary practical threshold is a 5% relative reduction. Experiment 3 uses counts of architectures with favourable paired effects only as screening evidence. State-resolved contrasts (Sec. 6.10) are post-hoc: they decompose published aggregate contrasts with the same comparators, use Holm correction within the states or groups of each intervention and metric, and are labelled exploratory (ten seeds) or descriptive (three seeds). Experiment 4 is a prospective protocol and will use fresh paired seeds and held-out operating points; its confirmatory family will be frozen before execution and will use the same separation between statistical significance and the 5% practical effect threshold.

6 Results

This section reports the empirical findings without forcing unsupported hypotheses into binary decisions. Formal hypothesis language is retained only for the two trajectory-versus-physical-fidelity hypotheses that satisfied the predefined support criteria; the remaining results are presented directly as effect estimates, secondary analyses, or descriptive screening evidence.

6.1 Experiment 1: retrospective validation-point selection

Experiment 1 is retained as retrospective evidence rather than as a confirmatory hypothesis test. The regenerated results are identical to the archived analysis, but the held-out outcomes had already been inspected before this completion run. Accordingly, the value of this experiment is descriptive: it shows how strongly different reference-only selectors concentrate learned-model failure regions under a reduced validation budget.

Table 1 reports the held-out recall results at all three budgets for which the archived selector table preserves exact values: 2%, 5%, and 10%. The budget dependence is substantial. At 2%, the frozen combined selector C_{top3} raises macro recall from 0.0201 under random sampling to 0.0679, a factor of approximately 3.38×. At 5%, the same

Table 1 Experiment 1 held-out validation recall across the archived 2%, 5%, and 10% budgets. Each budget reports macro recall and Recall-any. Bold values identify the largest recall within that budget and metric.

	2% budget		5% budget		10% budget	
Selector	Recall macro	Recall any	Recall macro	Recall any	Recall macro	Recall
random	0.0201	0.0201	0.0504	0.0504	0.1011	0.1
fixed_point	0.0222	0.0231	0.0487	0.0516	0.1052	0.1
trajectory_density	0.0207	0.0186	0.0528	0.0448	0.1033	0.1
I_J	0.0450	0.0350	0.0567	0.0499	0.0828	0.0
best_single (symmetric growth)	0.0652	0.0486	0.0696	0.0547	**0.4338**	**0.2**
best_combined (C_{top3})	**0.0679**	**0.0515**	**0.0770**	**0.0640**	0.4195	0.2

selector remains strongest, but the enrichment is smaller: 0.0770 versus 0.0504, or approximately 1.53×.

The most pronounced result occurs at the 10% validation budget. Here, symmetric growth reaches a macro recall of 0.4338 and C_{top3} reaches 0.4195, whereas random sampling reaches only 0.1011. Relative to random, these values correspond to enrichment factors of approximately 4.29× and 4.15×, respectively. In absolute terms, symmetric growth recovers 33.27 percentage points more of the critical set than random sampling, while the combined selector recovers 31.84 percentage points more. Thus, the 10% result is substantially stronger than the 5% result and is the clearest evidence in Experiment 1 that a reference-informed ranking can concentrate validation effort.

The ordering also changes with budget. C_{top3} is slightly stronger at 2% and 5%, whereas the single symmetric-growth selector becomes slightly stronger at 10%. The I_J selector provides another example of budget dependence: it exceeds random macro recall at 2% and 5%, but falls below random at 10% (0.0828 versus 0.1011). Consequently, Experiment 1 does not support a universal reference metric; it supports the narrower conclusion that reference dynamics can strongly enrich critical-region coverage in specific budget regimes.

The broader Experiment 1 protocol also evaluated whether an 80% recall target could be reached within a measured validation budget of up to 20%. That target was not reached. The 10% result should therefore be interpreted as strong concentration of failures relative to random sampling, not as evidence that a small validation subset captures most of the critical operating region.

The retrospective result is therefore most useful as a mechanism-discovery study. It demonstrates that a restricted validation budget can be allocated far more effectively than random sampling in some regimes—most clearly at 10%—while preserving the qualification that the effect is retrospective, budget dependent, and insufficient to establish a universal reduced-validation guarantee.

6.2 Experiment 2: supported trajectory-versus-fidelity hypotheses

Two predeclared hypotheses met their support criteria in the fresh-seed training experiment. The first states that low trajectory error can coexist with high Jacobian

Table 2 Experiment 2 primary observational-bias result: 14-state model, 5% training budget.

Metric	Physics–agnostic	95% CI	Rel. reduction	Holm p
$E_{A,\mathrm{rel}}$	-0.0296	$[-0.0324, -0.0269]$	2.09%	0.0039
E_{Φ}	-0.342	$[-0.681, -0.062]$	5.80%	0.064

error; the second states that low trajectory error can coexist with high frozen-local propagator error.

For the Jacobian hypothesis, the joint fraction of low-trajectory/high-$E_{A,\mathrm{rel}}$ cases was 0.029, above the predefined 0.01 threshold, with seed-level Spearman correlation 0.30. For the propagator hypothesis, the corresponding joint fraction was 0.031 and the seed-level Spearman correlation was 0.32. These results directly support the paper's multi-scale validation premise: trajectory accuracy and physical fidelity are related but not interchangeable. This interpretation is consistent with the system-identification literature, which distinguishes predictive fit from recovery of dynamical structure [8, 9], and with physics-informed learning studies showing that optimization success need not imply balanced learning of all physical objectives [3, 4].

6.3 Experiment 2: primary observational-bias effect

The primary observational-bias comparison is shown in Table 2. Negative physics-minus-agnostic differences favour physics-inspired training.

The Jacobian improvement is statistically significant after correction, but its 2.09% relative reduction is below the preregistered 5% practical threshold. The propagator improvement exceeds the practical threshold, but its corrected $p = 0.064$ does not meet the significance criterion. The appropriate interpretation is therefore effect-specific: physics-inspired sampling produces a statistically reliable but practically small Jacobian improvement, while the propagator effect is practically larger but statistically less certain after multiple-testing correction. This pattern is consistent with prior work showing that nonuniform sampling can materially change scientific-ML training while not guaranteeing uniform improvement across objectives [6, 7].

As a secondary distributional check, the upper-tail Jacobian error decreased by 1.9% with corrected $p = 0.004$, whereas the upper-tail propagator error decreased by 6.8% with $p = 0.084$. These tail effects reinforce the same mechanism-dependent interpretation and are not promoted to separate hypotheses.

6.4 Experiment 2: dependence on model complexity

The effect of observational bias changes materially with state dimension. Table 3 shows the 5% budget results.

The 3-state model shows the largest joint benefit in this table. The 7-state model is the clearest counterexample to a universal sampling claim: physics-inspired selection increases Jacobian error by 3.91%. The 14-state effect is beneficial but modest for $E_{A,\mathrm{rel}}$. This dimensional dependence shows directly that the observational-bias effect is not invariant with model complexity.

Table 3 Experiment 2 effect across model complexity at a 5% training budget. Positive percentages denote error reduction; a negative percentage denotes higher error under physics-inspired selection.

States	$E_{A,\mathrm{rel}}$ reduction	p	E_Φ reduction	p
3	4.50%	0.029	9.74%	0.016
7	−3.91%	0.012	2.82%	0.160
14	2.09%	0.012	5.80%	0.129

Table 4 Exploratory per-strategy results for Experiment 2, 14-state model at 5%. Lower is better. E_Φ is also shown using the per-fit pointwise median to reduce sensitivity to rare propagator blow-ups.

Strategy	E_A	median-based E_Φ
tire_transition	**1.178**	3.56
non_normality	1.220	4.30
finite_time / FTLE	1.288	3.78
random	1.298	4.34
I_J	1.312	4.22
physics_composite	1.338	4.03
physics_plus_coverage	1.339	3.68
deterministic_coverage	1.345	**3.61**
trajectory_density	1.601	4.96

6.5 Experiment 2: held-out manoeuvres

Physics-inspired sampling reduces $E_{A,\mathrm{rel}}$ in every held-out manoeuvre: 3.64% for crosswind, 1.83% for double lane change, 1.60% for low-friction cornering, 2.23% for sine dwell, and 2.11% for split-μ braking. Each reported comparison has $p = 0.020$.

The frozen-local propagator result is less uniform. Only crosswind and sine dwell show significant improvement ($p = 0.020$). Double lane change, low-friction cornering, and split-μ braking show relative changes of −1.96%, −7.62%, and −1.78%, respectively, i.e., higher propagator error with physics-inspired selection. Local Jacobian fidelity therefore generalizes more consistently across manoeuvres than the frozen-local finite-time metric.

6.6 Experiment 2: which training strategies actually lead?

The frozen family-level comparison is the confirmatory analysis, but the saved held-out fits permit a useful exploratory decomposition of the 14-state, 5% primary cell. Table 4 shows the per-strategy means over ten seeds. Because these rankings were not preregistered as a separate confirmatory family, they are interpreted descriptively.

Three observations follow. First, `tire_transition`, `non_normality`, and `finite_time` are the leading Jacobian selectors at 14 states across the 2%, 5%, and 10% budgets; the combined physics-plus-coverage policy is not the best Jacobian-training strategy. Second, in the primary 5% cell, physics-plus-coverage is essentially equal to physics-composite for E_A, worse than random by 3.2%, and only 0.4% better than

Table 5 v6 adaptive point-selection results in the primary 14-state, 5% cell. Differences are strategy minus the frozen-alpha comparator; negative values are better.

Method	Metric	Mean diff.	Rel. improvement	Holm p
M1 Lagrangian	$E_{A,\mathrm{rel}}$	-0.220	16.5%	0.00781
M1 Lagrangian	E_{Φ}	$+1.210$	-27.6%	0.00781
M2 primal–dual	$E_{A,\mathrm{rel}}$	-0.0339	2.55%	0.00781
M2 primal–dual	E_{Φ}	$+0.375$	-8.56%	0.00781

deterministic coverage. Third, for E_{Φ}, physics-plus- coverage improves over physics-composite by 10.5% and over random by 19.4%, but is statistically indistinguishable from deterministic coverage (Holm-adjusted $p = 0.85$). At 10%, the same pattern persists: its propagator error improves over physics-composite and random, but not over deterministic coverage (Holm-adjusted $p = 0.70$). Thus the propagator advantage of the mixed policy is attributable primarily to its 75% coverage component, not to a general benefit from adding physics-selected points.

The exploratory decomposition therefore sharpens the family-level result: Experiment 2 supports an effect of the broader physics-strategy family, but it does not support the stronger claim that combining physical metrics, or adding coverage to a physics composite, is optimal for training.

6.7 Experiment 2b: adaptive coverage-constrained selection and optimizer results

The v6 extension changes the interpretation of the earlier fixed-alpha experiment in an important way. The previous negative result applies to the specific union rule in which a fixed fraction of the budget is assigned to high-score physics points and the remainder to deterministic coverage. It does not imply that a physics-heavy training set is intrinsically harmful if coverage is enforced explicitly during selection.

Table 5 reports the two adaptive point-selection methods in the primary 14-state, 5% cell. The coverage-constrained Lagrangian selector M1 produces a large reduction in Jacobian error relative to the frozen-alpha comparator: the mean paired difference is -0.220, with 95% CI $[-0.222, -0.218]$, corresponding to a 16.5% relative improvement and Holm-adjusted $p = 0.00781$. The same selection, however, increases propagator error by 27.6% ($+1.21$, 95% CI $[0.818, 1.61]$, Holm-adjusted $p = 0.00781$). The adaptive point-selection result is therefore mechanism-specific: M1 substantially improves local Jacobian fidelity but degrades frozen-local finite-time fidelity.

Table 5 isolates the point-selection axis so that the M1/M2 confidence intervals and interpretation remain visible. The same M1/M2 outcomes are repeated in Panel A of Table 7, where the v6 results are separated by matched baseline comparator to avoid treating unlike contrasts as a single ranking.

M2 exhibits the same directional trade-off but with smaller magnitude. Its Jacobian error decreases by 2.55% (-0.0339, 95% CI $[-0.0379, -0.0304]$), which is statistically resolved after correction but below the 5% practical threshold, while propagator error increases by 8.56% ($+0.375$, 95% CI $[0.252, 0.511]$). Because M2 also contains the

Table 6 Primary-cell v6 control weights and realized point-set composition. The control/weight quantity has a method-specific meaning and must not be confused with the realized PHY fraction $f_{\mathrm{PHY}} = |S_{\mathrm{PHY}}|/|S|$. For the loss-balancing methods, the selected point set is fixed, so f_{PHY} remains 0.262 while only the training-loss weight changes.

Method	Control / weight value	f_{PHY}
Frozen v5 physics+coverage	$\alpha_{\mathrm{base}} = 0.250$	0.262
M1 Lagrangian	$\alpha^{\star}_{\mathrm{M1}} = 0.973$	0.994
M2 primal–dual, last round	$\alpha_{\mathrm{M2,last}} = 0.00691$	0.00113
dualweight, final loss weight	0.186	0.262
lra, final loss weight	0.00831	0.262
gradnorm, final loss weight	0.249	0.262
softadapt, final loss weight	0.393	0.262
relobralo, final loss weight	0.288	0.262

reused-constraint limitation described in Sec. 5.3, its primary-cell behavior is not used to make a general recommendation.

The allocation values in Table 6 explain why M1 is not equivalent to the old fixed-fraction sweep. In the primary cell, the frozen v5 method uses the configured allocation $\alpha_{\mathrm{base}} = 0.25$, while the realized composition of the final set is $f_{\mathrm{PHY,base}} = 0.262$. These quantities must not be identified with each other. M1 selects the score-versus-coverage objective weight $\alpha^{\star}_{\mathrm{M1}} = 0.973$ and produces $f_{\mathrm{PHY,M1}} = 0.994$ while still satisfying the explicit normalized-coverage constraint. Thus nearly the complete selected set can be physics-labelled without losing the prescribed facility-location coverage. In contrast, the last-round M2 objective weight is $\alpha_{\mathrm{M2,last}} = 0.00691$, with a realized $f_{\mathrm{PHY,M2}} = 0.00113$ in this cell. The symbols in the second column of Table 6 are control or optimization weights with method-specific meanings; the third column alone reports the realized composition of the selected set.

Because the v6 interventions do not all use the same baseline, their results must be interpreted within matched comparator groups rather than as one global ranking. Table 7 therefore separates the primary results into three panels. Panel A isolates the effect of changing the selected training set: M1 and M2 are compared with the frozen v5 `physics_plus_coverage` selector configured with $\alpha_{\mathrm{base}} = 0.25$, whose realized primary-cell composition is $f_{\mathrm{PHY,base}} = 0.262$. Panel B isolates the effect of loss reweighting: all five adaptive weighting rules are compared with the unweighted Adam fit on exactly this same frozen point set and seed. Their adaptive quantity $\alpha_t = w_{\mathrm{phy}}/2$ is a loss weight and does not alter $f_{\mathrm{PHY,base}}$. Panel C isolates the optimizer effect: Adam followed by L-BFGS-B is compared with Adam on the same data and objective. Differences and relative improvements should therefore be compared directly only within a panel.

Panel A shows that changing the point-selection rule has the largest observed effect. Relative to the frozen v5 mixed selector, M1 improves $E_{A,\mathrm{rel}}$ by 16.5% but worsens E_{Φ} by 27.6%; M2 provides a smaller 2.55% Jacobian improvement and an 8.56% propagator deterioration. These are selection effects because the baseline is the previous frozen selection policy.

Panel B answers a different question. Here the point identities are held fixed and only the relative PHY/COV loss weights change. None of the five adaptive weighting

Table 7 v6 primary results separated by matched baseline comparator in the 14-state, 5% cell. Within every panel, the reported difference is strategy minus baseline, so negative values indicate lower error and positive relative improvement indicates improvement over that panel's baseline. Results from different panels should not be interpreted as a direct ranking because their baselines differ.

Method	Metric	Mean diff.	Rel. improvement	Holm p
Panel A — Baseline: frozen v5 selector ($\alpha_{\text{base}} = 0.25,\ f_{\text{PHY,base}} = 0.262$)				
M1 Lagrangian	$E_{A,\text{rel}}$	-0.220	16.5%	0.00781
M1 Lagrangian	E_Φ	$+1.210$	-27.6%	0.00781
M2 primal–dual	$E_{A,\text{rel}}$	-0.0339	2.55%	0.00781
M2 primal–dual	E_Φ	$+0.375$	-8.56%	0.00781
Panel B — Baseline: unweighted Adam fit on the same frozen-alpha point set and seed				
dualweight	$E_{A,\text{rel}}$	$+0.00618$	-0.465%	0.0195
dualweight	E_Φ	$+0.143$	-3.31%	0.193
lra	$E_{A,\text{rel}}$	$+0.00783$	-0.589%	0.0488
lra	E_Φ	$+0.371$	-8.48%	0.0195
gradnorm	$E_{A,\text{rel}}$	$+0.00298$	-0.224%	0.0781
gradnorm	E_Φ	$+0.00526$	-0.139%	0.922
softadapt	$E_{A,\text{rel}}$	$+0.0113$	-0.853%	0.0195
softadapt	E_Φ	$+0.417$	-9.60%	0.0352
relobralo	$E_{A,\text{rel}}$	$+0.00696$	-0.524%	0.0195
relobralo	E_Φ	$+0.201$	-4.62%	0.193
Panel C — Baseline: Adam on the same selected data and objective				
Adam–L-BFGS	$E_{A,\text{rel}}$	$+0.00618$	-0.465%	0.00391
Adam–L-BFGS	E_Φ	$+0.0570$	-1.26%	0.275

rules improves the mean Jacobian metric relative to its directly matched unweighted fit, and none provides a positive relative improvement in E_Φ. This negative result therefore concerns post-selection loss balancing rather than data selection.

Panel C isolates optimizer refinement. Adding full-batch L-BFGS-B after Adam slightly but significantly increases Jacobian error by 0.465% in the paired primary comparison, while its 1.26% propagator deterioration is not statistically resolved after correction. The optimizer comparison therefore provides no evidence that second-stage L-BFGS-B improves the two structural-fidelity outcomes considered here.

The v6 audit contains 2160 fits. No physics-agnostic fit fails the learner gate under any optimizer, whereas 39 strategy-induced failures are retained as evidence rather than removed. In addition, all 360 Adam refits of the v5 strategies are bit-identical to the corresponding v5 runs. This audit confirms that the new results arise from the added selection, weighting, and optimizer axes rather than from an unintended change to the frozen Adam baseline.

The physical, state-level interpretation of these aggregate results—which vehicle states and couplings each selection, weighting, and optimizer intervention changes—is given together with the prior and architecture interventions in Sec. 6.10.

Table 8 Experiment 3 screening counts: architectures, out of 14, for which the paired effect lowers error. E_A uses the mean-over-points summary used by the frozen analysis; E_Φ uses the recommended median-over-points summary to avoid domination by rare exponential blow-ups.

Budget	Metric	Observational	Inductive $I1$	Interaction
2%	E_A	0	3	11
2%	median E_Φ	9	3	10
10%	E_A	1	3	7
10%	median E_Φ	7	9	11

6.8 Experiment 3: architecture screening

All 336 screening fits passed the learner-quality gate. Across the eight balanced observational × inductive × budget cells, mean Jacobian error ranges from 0.712 for `medium256` and 0.768 for `plain64` to 1.008 for `deep128`. This spread requires a more careful interpretation than a simple architecture ranking.

Most architectures remain close to the structural fidelity of their fitted affine output model. At 14 states, an affine-only predictor gives mean E_A values of approximately 1.02–1.05, whereas most skip architectures remain in the 0.92–1.01 range. The best possible constant Jacobian (the mean reference Jacobian over the evaluation set) has mean $E_A = 0.906$. Two specific cases account for most of the apparent architecture advantage: `plain64`, which removes the affine skip and must learn the complete field, and `medium256` at the 10% budget, where sufficient capacity and data produce much lower Jacobian error. At 2%, even `medium256` remains near $E_A \approx 1$. Activation, depth, and moderate-width changes do not produce a smooth monotonic architecture trend.

This result also clarifies why a low vector-field training loss is not enough: several deep/high-capacity models achieve very small final loss while retaining poor Jacobian fidelity. The network can improve nominal residual prediction without learning the required local derivative structure.

6.9 Experiment 3: observational, inductive, and interaction effects

For Jacobian error, the mean-based screening contrasts show a strongly negative observational effect: physics-plus-coverage lowers mean E_A in 0 of 14 architectures at 2% and only 1 of 14 at 10%. The analytical prior lowers mean E_A in 3 of 14 architectures at each budget. The interaction contrast is positive in 11 of 14 architectures at 2% and 7 of 14 at 10%, but this does not imply that the combined cell is good: a positive interaction can arise because the combination is merely less harmful than the sum of two harmful main effects.

Finite-time error requires a different summary. Mean E_Φ is dominated by rare points at which a learned Jacobian has a spurious positive real-part eigenvalue and the matrix exponential explodes. In the largest tanh networks, the worst single-point E_Φ reaches 10^{11}–10^{12}, even though the typical per-fit values remain order one. We therefore use the pointwise median E_Φ for the architecture screening and retain extreme quantiles or maxima as separate instability indicators.

The median propagator result is more balanced than the Jacobian result. At 10%, the analytical prior improves typical E_Φ in 9 of 14 architectures and the interaction is favourable in 11 of 14. The state-resolved analysis of Sec. 6.10 identifies the mechanism more precisely than the architecture counts. With an affine skip, the attitude and heave kinematic rows are already reproduced exactly under $I0$, so the operative part of the prior is the rotating-frame bilinear coupling, which the $I0$ learner largely misses and the $I1$ learner partly cancels. Without the affine skip, the prior also restores the exact kinematic rows and produces the largest finite-time benefit. In both cases the affected entries are a small fraction of a Jacobian dominated by tyre and other nonlinear terms, which is why the effect is more visible in the propagator than in the global Frobenius Jacobian error.

The Experiment 3 results in Secs. 6.8–6.9 and Table 8 are consequently not evidence that one architecture or bias is universally best. Rather, it shows three specific phenomena: the tested mixed sampling policy is consistently harmful for mean Jacobian fidelity; the analytical prior has a small Jacobian effect but a more frequent typical finite-time benefit at 10%; and architecture capacity matters only in specific regimes, especially when enough data are available.

6.10 Physical effect of each training intervention

The preceding subsections report *whether* each intervention changes the global fidelity metrics. This subsection reports *which part of the vehicle physics* changes, at two levels. Level 1 is the global effect on nominal, trajectory, Jacobian, propagator, and stability metrics (Table 11). Level 2 is the effect on individual state equations, state influences, perturbation directions, couplings, and local modes (Table 10, Figs. 1 and A1). Table 9 summarizes what each intervention changes in training and what it changes physically.

Scope, provenance, and evidential status.

The analysis is a read-only recomputation from frozen checkpoints. No network was retrained, no point selection was rerun, and no frozen result was modified. It covers 716 saved fits: the 160 Experiment 2 (v5) fits of the 14-state, 5% cell on their own held-out set; the 220 v6 fits of the same cell under Adam and Adam→L-BFGS-B on the v6 held-out set (2000 points per fit); and all 336 Experiment 3 fits on their 600 shared development audit points. For every fit the saved $E_{A,\mathrm{rel}}$, E_Φ, and E_F were reproduced to within 9.1×10^{-13} before any new quantity was kept. Checkpoint SHA-256 and parameter hashes were unchanged, a guard blocked every training and selection entry point (none was called), and the file trees of the three frozen experiments (30,703 files) were identical before and after. Each state-level contrast decomposes a published aggregate contrast and uses the same comparator. Paired seed differences are tested with the exact Wilcoxon signed-rank test, Holm-corrected within the 14 states (or seven groups) of each intervention and metric, with a 2000-resample seed bootstrap. All v5/v6 state-level results are *post-hoc exploratory*. All Experiment 3 results (three seeds) are *descriptive*, and none can be statistically resolved. The symbol rule of Table 10 was fixed before any contrast was computed. One amendment followed the first smoke test, before any text was written: two relative errors that are both below 1% are marked

as exactly represented. Experiment 4 is a planned protocol in this manuscript, so no state-level effect is reported for it.

A common structural fact.

Before the individual interventions, one property of the 14-state learners conditions every result below. The reference rows of ϕ, θ_p, z, $\dot{z}$, and F_{wind} are constant over the domain. The fitted affine output map therefore reproduces them exactly, and no intervention that keeps the affine skip changes them (symbol 0^*). The physically informative equations are the nine state-dependent rows: v_x, v_y, r, $\dot{\phi}$, $\dot{\theta}_p$, and the four wheel speeds. In the baseline tanh learner only two of them are represented as genuinely nonlinear (Table 13). The lateral (v_y) and roll-rate ($\dot{\phi}$) rows reach $V_i = 0.66$–0.70, against reference values of 0.68–0.70, with neural shares of 0.68–0.86. The longitudinal, pitch-rate, and wheel rows stay nearly affine ($V_i = 0.04$–0.15 against reference values of 0.59–0.81), and the yaw row is intermediate ($V_r \approx 0.32$). Any intervention that leaves the architecture unchanged can therefore alter the longitudinal, pitch-rate, and wheel equations mainly through the ridge-fitted affine map, which is refitted on the selected points. It can alter the lateral and roll equations through the neural residual.

6.10.1 A. Random and physics-agnostic training

What changes in training. Random sampling draws operating points without any physical criterion. Deterministic coverage spreads the same budget evenly across manoeuvres and trajectory progression. Random sampling is the comparator for interventions B and C. The coverage-versus-random contrast isolates the effect of geometric support without physics information.

Global effect. Relative to random sampling (v6 held-out), deterministic coverage leaves $E_{A,\text{rel}}$ essentially unchanged (+3.6%). It lowers E_Φ by 19.9% (median 15.8%) and the spectral-abscissa error by 35.4% (both resolved).

State/subsystem effect. No state equation improves. Two wheel rows deteriorate slightly (ω_{FR} +6%, ω_{RR} +10%, resolved). The finite-time response to yaw, roll, roll-rate, heave, and wheel-speed perturbations improves by 13–37% (mean; resolved for most states). Under random sampling, roll and roll-rate perturbations are the worst-propagated directions (median $E_{\Phi,\text{dir}} \approx 2.7$–$3.0$), followed by yaw (1.6) and lateral velocity (1.2). The wheel equations have the largest row errors.

Physical meaning. Broad support of the domain does not make any local equation more accurate. Its benefit is a better-conditioned learned linearization: spectral-abscissa error −35% (resolved), with fewer spurious growth directions and better stability-class preservation (both not resolved). This appears as better finite-time propagation of body-attitude and wheel perturbations.

Evidence. Exploratory; ten paired seeds on the v6 held-out set.

6.10.2 B. Physics-inspired point selection (Experiment 2 family)

What changes in training. The selected operating points: each of eleven reference-metric selectors observes a different set of high-score regimes. The family effect is the per-seed

median over the twelve physics labels minus the mean over the three agnostic labels, exactly as in the confirmatory analysis.

Global effect. The confirmatory result is unchanged: $E_{A,\mathrm{rel}}$ improves by 2.09% (resolved but below the practical threshold) and E_Φ by 5.80% (not resolved after Holm correction; Table 2). The state-level recomputation adds exploratory global outcomes: E_F −14.3%, spectral-abscissa error −15.8%, stability-class preservation +15.4%, and long-horizon trajectory RMSE −22.6%, all resolved.

State/subsystem effect. Only one state equation changes by more than 5%: the lateral-velocity row improves by 10.3% (0.713→0.640, resolved). The finite-time response changes far more. Median directional errors fall by 35.1% for yaw-rate perturbations (1.69→1.09), 39.8% for roll (3.51→2.10), 36.2% for roll rate, 27.2% for pitch, 12–20% for three of the four wheel speeds, and 45.3% for the crosswind state; heave is unchanged. Nominally, the longitudinal (−21%), yaw (−15%), pitch-rate (−25%), and wheel derivatives become more accurate. The influence of lateral velocity on the other equations (column v_y) becomes 29% worse.

Physical meaning. Observing physically informative regimes barely changes the local differential equations; the 2.09% global $E_{A,\mathrm{rel}}$ gain is a lateral-row effect. What it does change is the finite-time propagation of yaw, roll, and pitch perturbations, together with local stability. The pointwise-median view thus resolves a body-attitude benefit that the mean E_Φ, dominated by rare blow-ups, does not.

Evidence. Exploratory decomposition of a confirmatory family contrast. The family median mixes eleven distinct selections, so no state effect is attributed to any single selector.

6.10.3 C. Frozen physics-plus-coverage selection

What changes in training. The observations: 25% of the budget goes to diversity-constrained physics points and 75% to deterministic coverage ($f_{\mathrm{PHY}} = 0.262$).

Global effect. Against random sampling: E_Φ −18.7%, spectral-abscissa error −35.4% (resolved), and $E_{A,\mathrm{rel}}$ unchanged. Against deterministic coverage, every global metric changes by less than 5% except trajectory RMSE (+53.8%, not resolved).

State/subsystem effect. Against random sampling the yaw row worsens by 8% and three wheel rows by 6–8% (resolved), while the roll-rate row improves by 6%. Perturbation propagation improves for every state, by 7–35% (median, resolved for all 14). Against deterministic coverage, no row changes and only two directional improvements are resolved: pitch rate (−15%) and the crosswind state (−28%). In Experiment 3, with the A0 learner at 10%, the same policy worsens the longitudinal, pitch-rate, and wheel rows by 11–13% (descriptive).

Physical meaning. The finite-time benefit of the mixed policy is the coverage benefit of intervention A. The 25% physics component adds only a better pitch-rate and disturbance response and slightly degrades the yaw and wheel equations. This is the state-level counterpart of the per-strategy result in Sec. 6.6.

Evidence. Exploratory (v6); descriptive (Experiment 3).

6.10.4 D. M1 coverage-constrained Lagrangian selection

What changes in training. The observations: the selected set becomes 99.4% physics-labelled while retaining 90% of the facility-location coverage.

Global effect. $E_{A,\mathrm{rel}}$ −16.5%, E_Φ +27.6% (median +22.4%), spectral-abscissa error +63.6%, and stability-class preservation 76.9%→63.5% (all resolved); E_F +9.9% (not resolved).

State/subsystem effect. M1 makes the longitudinal (−28.0%), pitch-rate (−27.8%), and three wheel equations (−19 to −23%) more accurate. It makes the lateral-velocity (+44.4%, 0.62→0.90) and roll-rate (+63.6%, 0.65→1.06) equations less accurate; all of these are resolved, and the yaw row is unchanged. The influence of lateral velocity (column v_y, +78%) and of roll rate (column $\dot{\phi}$, +40%) on the rest of the vehicle also degrades. In block terms this reproduces the frozen analysis exactly (Table A2). The dominant blocks pitch←longitudinal (−10.5%), wheels←longitudinal (−6.3%), pitch←wheels (−6.9%), and wheels←wheels (−2.3%) improve and carry about 90% of the baseline squared discrepancy. The smaller roll and lateral blocks deteriorate by 76–152%. Perturbation propagation worsens for longitudinal (+20.5%), lateral (+23.5%), pitch (+24.3%), and pitch-rate (+37.0%) perturbations (median, resolved), and for roll-rate perturbations (+47% mean, resolved; +12% median, not resolved). The pointwise change of E_Φ is associated with the change in lateral ($\rho = 0.84$) and roll ($\rho = 0.72$) directional error. Matched modes dominated by lateral and yaw coordinates lose fidelity: damping-ratio errors increase by 126% and 146% (resolved) and natural-frequency errors by 250% and 181% (Table A1).

Physical meaning. The decomposition identifies the route. Table 13 and the affine-only Jacobian show that M1's longitudinal, pitch-rate, and wheel gains are already present in the affine component refitted on the M1 points (v_x row 2.92→2.10, $\dot{\theta}_p$ row 2.50→1.80). Concentrating observations on tyre-slip-dominated, high-score states therefore improves the *average slopes* of the stiff longitudinal-slip, pitch-rate, and wheel-spin equations, which dominate the Frobenius norm. The same concentration starves the lateral-force and roll-moment equations that the neural residual represents. Those smaller equations govern the slow lateral, yaw, and roll modes and therefore the finite-time response. M1 is a clear instance of a training process that lowers a global structural metric while degrading the states and couplings most relevant to perturbation propagation.

The degradation is distributed rather than localized. It is not concentrated in any predefined reference regime: for example, high reference non-normality holds 14% of the points but only 9% of the summed positive ΔE_Φ. ΔE_Φ is associated with the *learned* spectral abscissa ($\rho = 0.42$) and the learned amplification $\sigma_{\max}(\widehat{\mathbf{\Phi}})$ ($\rho = 0.68$), not with the reference spectral abscissa ($\rho = 0.02$). For 53.5% of the held-out points (seed means), $E_{A,\mathrm{rel}}$ improves while E_Φ worsens, and the pointwise changes $\Delta E_{A,\mathrm{rel}}$ and ΔE_Φ are nearly uncorrelated ($\rho = 0.03$).

Evidence. Exploratory; ten seeds; the block and modal values reproduce the frozen post-hoc analysis.

6.10.5 E. M2 sequential primal–dual selection

What changes in training. The observations, chosen sequentially: in the primary cell, the reused-constraint limitation drives the multiplier to its clip, and the final set is almost pure coverage ($f_{\mathrm{PHY}} = 0.001$).

Global effect. $E_{A,\mathrm{rel}}$ −2.6% (below the practical threshold), E_Φ +8.6% (median +11.3%), spectral-abscissa error +30.5%, and spurious-growth rate +63% (resolved).

State/subsystem effect. M2 was tested rather than assumed to be a weaker M1, and its signature differs. It does *not* reproduce M1's longitudinal, pitch-rate, and wheel row improvements; those rows change by less than 5% except ω_{FL} (−8.5%). Nor does it reproduce M1's lateral and roll-rate row degradation (v_y +1.5%, $\dot{\phi}$ +5.0%, neither resolved). Instead the yaw row improves by 7.9% (resolved). The lateral-velocity column improves by 17.6% (M1: +78%), and the block lateral←longitudinal improves by 16.9% (M1: +100%). The roll-rate and pitch columns, by contrast, worsen by 32% and 58%. Its finite-time degradation affects other directions than M1's: yaw (+14.8%), roll (+21.6%), roll rate (+23.6%), two wheel speeds (+20, +29%), and the crosswind state (+58.6%), median and resolved. Longitudinal and lateral perturbations are unaffected.

Physical meaning. M2 reproduces M1's aggregate $E_{A,\mathrm{rel}}$-better/E_Φ-worse pattern through a different mechanism. M1 changes the *equations* of the stiff and of the lateral subsystems. M2 leaves the equations nearly intact but changes how roll rate and pitch attitude *influence* the remaining states, adding spurious growth directions that degrade yaw, roll, and disturbance propagation.

Evidence. Exploratory. Because the M2 set is shaped by the known reused-constraint limitation, this signature characterizes the implemented selector, not the ideal primal–dual method.

6.10.6 F. Adaptive loss weighting

What changes in training. The data stay fixed; the relative weight of the physics-labelled and coverage-labelled residuals in the parameter updates changes. The final PHY weight fractions are 0.186 (DualWeight), 0.0083 (LRA), 0.249 (GradNorm), 0.393 (SoftAdapt), and 0.288 (ReLoBRaLo).

Global effect. No balancer improves a matched global mean. LRA worsens E_Φ by 8.5% (median 11.5%, resolved), and the others change $E_{A,\mathrm{rel}}$ and E_Φ by less than 5% or without resolution.

State/subsystem effect. The affine component is identical across balancers because the point set is identical, so every difference comes from the neural residual and therefore from its lateral/roll subspace. Reweighting does not preferentially *improve* any subsystem: across the state equations, state influences, nominal derivatives, and perturbation directions there is not a single resolved improvement of 5% or more under any balancer. DualWeight, GradNorm, and ReLoBRaLo change every row by less than 5%. Their finite-time changes are at most 17% and unresolved, except that ReLoBRaLo resolves 16–29% increases in the natural-frequency error of yaw-, roll-, and pitch-dominated modes. SoftAdapt degrades roll and roll-rate propagation by about 23% (not resolved). LRA, which almost ignores the physics-labelled residuals, worsens the

lateral-velocity (+13.7%) and roll-rate (+12.1%) equations (resolved). It degrades every perturbation direction (yaw +40.6%, roll +40.7%, crosswind +98%, median, resolved).

Physical meaning. Under the present affine-skip learner, the measurable effects of reweighting physics-versus-coverage residuals are concentrated in the lateral/roll residual subspace. Other equations receive neural corrections too small to produce a resolved state-level improvement. The measured effects are either negligible or degrading, and the clearest degradation occurs where the physics residuals are effectively switched off. No balancer improves roll, yaw, or any other subsystem.

Evidence. Exploratory; ten seeds.

6.10.7 G. Optimizer refinement: Adam → L-BFGS-B

What changes in training. The same data and objective are minimized with a second-stage quasi-Newton method.

Global effect. $E_{A,\mathrm{rel}}$ +0.5% (resolved but negligible), E_Φ unchanged, nominal E_F +18.9%, and spectral-abscissa error +16.0% (resolved; median over the 11 strategies, the O1 form).

State/subsystem effect. No state equation changes; every row changes by less than 5%. The optimizer nevertheless redistributes error in a reproducible way. The neural share of the near-affine rows roughly triples (e.g. wheel rows 0.04–0.06→0.12–0.21). The columns of the linear attitude, heave, and disturbance states degrade by 65–162% (e.g. θ_p 0.28→0.72, resolved): the refined network makes the nonlinear equations spuriously sensitive to ϕ, θ_p, z, $\dot{\theta}_p$, and F_{wind}. Finite-time propagation improves for yaw (−14.6%), roll (−21.4%), and roll-rate (−12%) perturbations but worsens for pitch (+21.4%), pitch rate (+16.1%), heave (+22.9%), longitudinal (+7.7%), and crosswind (+24%) perturbations (median, resolved). Three of the four nominal wheel-speed derivatives deteriorate by 26–91% (resolved), and matched-mode frequency errors increase by 77–310%.

Physical meaning. Second-order refinement spends additional nonlinear capacity on spurious couplings, not on the state dependence missing from the longitudinal, pitch-rate, and wheel equations. It yields no improvement of any state equation, and its finite-time gains for yaw and roll perturbations are offset by losses for pitch, heave, and disturbance perturbations.

Evidence. Exploratory. The v6 L-BFGS-B fits are not bit-reproducible across BLAS thread counts; the analysis evaluates the frozen checkpoints themselves, so it is deterministic for those fits.

6.10.8 H. Analytical prior $I1$ versus $I0$

What changes in training. The parameterization: exact rotating-frame terms enter the v_x and v_y equations, and kinematic integrators enter the ϕ, θ_p, and z equations (Eqs. (34)–(35)). The loss is unchanged.

Global effect. For A0 the effect on $E_{A,\mathrm{rel}}$ is below 1% at both budgets. At 10% the median E_Φ falls by 6.3%; at 2% the mean E_Φ rises by 13.4% (descriptive). Without the affine skip (plain64, 10%), E_Φ falls by 25.9%.

State/subsystem effect. Table 12 tests the equations into which the prior is inserted. With an affine skip, the kinematic rows ϕ, θ_p, and z are already exact under $I0$, so the kinematic part of the prior is redundant there. The rotating-frame terms are the only part that matters. The $I0$ learner lacks 55% of $\partial\dot{v}_x/\partial r$ and of $\partial\dot{v}_y/\partial v_x$ at 10% (86% and 132% at 2%), since a bilinear term has no constant slope that the affine map could capture. $I1$ supplies them: the missing fractions fall to -0.03 and 0.24 at 10%. The learned residual, however, cancels 47% of the supplied $\partial\dot{v}_x/\partial v_y$ term, which $I0$ had already learned. Despite these entry-level corrections, the v_x row error is unchanged (1.01→1.01) and the v_y row improves by only 6.7%, because tyre forces dominate both rows. Effects do not transfer to the yaw, roll-rate, pitch-rate, or wheel equations; all change by less than 5%. Without the affine skip the picture changes. The prior restores the exactness of the ϕ, θ_p, and z rows (e.g. $0.228 \to 2 \times 10^{-5}$), improves the v_y row by 14.6%, and improves the propagation of lateral (-28.7%), yaw (-30.6%), and roll (-25.6%) perturbations. The $\dot{z}$ row, which receives no prior, worsens by 17%.

Physical meaning. The prior does not necessarily improve the equations that it encodes. Its effect is conditioned by the architecture. When an affine skip already represents the linear kinematics, only the small rotating-frame entries remain to be supplied, and they are partly cancelled. When the network must learn the whole field, the prior removes a representation burden and the benefit appears in the lateral, yaw, and roll finite-time response. The more frequent median-E_Φ benefit of $I1$ at 10% (Table 8) should therefore be attributed to the rotating-frame couplings and, for no-skip learners, to the kinematics. It should not be attributed to the kinematics in skip learners.

Evidence. Descriptive; three seeds.

6.10.9 I. Architecture: affine skip, capacity, depth, and activation

What changes in training. The representable tangent map: which rows can become state dependent, and how strongly.

Global effect. Relative to A0 at 10%: medium256 lowers $E_{A,\mathrm{rel}}$ by 57.7% and median E_Φ by 14.2%, but its mean E_Φ explodes, and the spectral-abscissa error rises by 566%. plain64 lowers $E_{A,\mathrm{rel}}$ by 10.2% but raises E_Φ by 35.7%. Depth 6 and ReLU raise E_Φ by 30% and 41%, and width 256 lowers it by 13%. At 2%, medium256 is within a few per cent of A0.

State/subsystem effect. The affine-versus-neural decomposition (Table 13) explains these differences row by row. At 10%, medium256 makes the longitudinal, yaw, pitch-rate, and wheel rows genuinely state dependent ($V_i = 0.51$–0.71, close to the reference 0.59–0.75). Their errors fall by 47–70% (e.g. $\dot{\theta}_p$ 0.95→0.40, ω_{RL} 5.15→1.56), and the lateral and roll-rate rows improve by 14% and 21%. It improves body and wheel dynamics together. At 2% the same network remains affine in those rows ($V_i = 0.08$–0.16) and no row changes. Removing the affine skip (plain64) makes every row neural. The longitudinal, pitch-rate, and wheel rows improve by 11–20%, but the five linear rows lose their exactness (e.g. ϕ row error $1 \times 10^{-4} \to 0.23$, F_{wind} row 0.87). The lateral and roll-rate rows worsen by 9%, and every perturbation direction is propagated 25–109% worse. Depth 6 degrades the lateral (+16%) and yaw (+13%) rows. ReLU degrades the lateral (+42%) and roll-rate (+38%) rows and every perturbation direction

(+33 to +185%). Width 256 improves the ω_{FL} row (−16%) and most perturbation directions (7–29%).

Physical meaning. This is the physical mechanism behind "architecture conditions sampling effects". With the baseline learner, the data can influence the longitudinal, pitch-rate, and wheel equations only through average slopes, by way of the refitted affine map (M1's route), and the lateral and roll equations through a neural residual (the balancers' route). Large capacity with enough data is the only screened change that makes the stiff longitudinal-slip and wheel equations state dependent. Removing the affine skip trades exact linear kinematics for nonlinear slopes and is harmful at finite time. Larger networks are not uniformly more physical: medium256 improves the local equations but introduces rare spurious unstable modes.

Evidence. Descriptive; three seeds per cell.

6.10.10 J. Experiment 4 optimal design of experiments

Experiment 4 is defined in Sec. 5.6 as a planned protocol, and no Experiment 4 result is integrated in this manuscript. Neither aggregate nor state/subsystem-level attribution is reported for it.

Validation-metric correspondence at state level.

The three correspondence levels of Eq. (31) extend naturally to a state-resolved training effect. Write $\Delta^{(j)}_{m\to k}$ for the effect of training process m on the error mechanism k associated with state or group j, where k is a row, column, or directional error. The symbols in Table 10 are exactly these quantities with their evidential status. The table shows that $\Delta_{m\to k}$ and $\Delta^{(j)}_{m\to k}$ can differ in sign. For M1, $\Delta_{\mathrm{M1}\to E_{A,\mathrm{rel}}} < 0$ while $\Delta^{(v_y)}_{\mathrm{M1}\to\mathrm{row}} > 0$ and $\Delta^{(\dot{\phi})}_{\mathrm{M1}\to\mathrm{row}} > 0$.

Table 9 What each training process changes in training and what it changes physically. Global changes are mean relative changes against the matched comparator (negative = lower error). "Resolved": Holm-adjusted exact Wilcoxon $p < 0.05$ over ten paired seeds; Experiment 3 entries (three seeds) are descriptive. State-level entries are post-hoc exploratory.

Training process (comparator)	What changes in training	$E_{A,\mathrm{rel}}$ effect	E_Φ effect	Dominant state-level effect	Interpretation
A Det. coverage (random)	which domain regions are observed, no physics	+3.6% (n.s.)	−19.9%	no equation improves; yaw, roll, heave, wheel perturbations propagate 13–37% better	support conditions the linearization, not the equations
B Physics family (agnostic; Exp. 2)	which physical regimes are observed	−2.1% (confirmatory)	−5.8% (n.s.)	only the v_y row improves (−10%); yaw, roll, pitch, wind perturbations −27 to −45% (median)	finite-time attitude benefit hidden by mean E_Φ
C Physics+coverage (random / det. cov.)	25 % physics + 75 % coverage	+3.2% / −0.4%	−18.7% / +1.6%	vs coverage only pitch-rate and wind propagation improve	benefit is the coverage component
D M1 (frozen P+C)	observations concentrated on high-score points under coverage constraint	−16.5%	+27.6%	v_x, $\dot{\theta}_p$, wheel rows −19 to −28%; v_y +44%, $\dot{\phi}$ +64%; lateral/yaw modes degrade	better average slopes of stiff subsystems; local derivative gain ≠ finite-time fidelity
E M2 (frozen P+C)	observations chosen sequentially (near-pure coverage here)	−2.6% (n.s. practical)	+8.6%	equations nearly unchanged (r −8%); roll-rate and pitch influences worsen; yaw, roll, wind perturbations +15 to +59%	same aggregate trade-off, different mechanism from M1
F Loss balancers (unweighted Adam)	weight of PHY vs COV residuals in the updates	+0.2 to +0.9%	+0.1 to +9.6%	no resolved improvement anywhere; LRA degrades v_y, $\dot{\phi}$ rows and all perturbation directions	measurable effects are concentrated in the neural lateral/roll subspace; negligible or harmful
G Adam→L-BFGS-B (Adam)	numerical minimizer of the same objective	+0.5%	+1.3% (n.s.)	no row changes; spurious sensitivity to $\phi, \theta_p, z, F_\mathrm{w}$ (+65–162%); yaw/roll	redistributes, does not improve, physical fidelity

Table 10 Training process × physical state/subsystem. Each cell gives two symbols: Jacobian-row error $E_{A,\mathrm{row}}$ (local differential equation of that state or group) / directional propagator error $E_{\Phi,\mathrm{dir}}$ (finite-time response to perturbing that state or group; pointwise median). ++/—: improvement/degradation $\geq 10\%$, +/−: 5–10%, both statistically resolved (Holm-adjusted exact Wilcoxon $p < 0.05$ within the 14 states or 7 groups of that intervention and metric, seed-bootstrap CI excluding 0); parentheses: $|\Delta| \geq 5\%$ but not resolved, or Experiment 3 (three seeds, always descriptive); 0: $|\Delta| < 5\%$; 0^*: both arms below 1% relative error (linear reference row reproduced exactly); $\pm^*$: exactness gained/lost. Comparators as in Table 9. All entries are post-hoc exploratory (v5/v6, ten seeds) or descriptive (Experiment 3).

Training process	v_x	v_y	r	roll	pitch	heave	wheels
A: det. coverage vs random	0/(+)	0/(+)	0/(++)	0/(+)	0/0	0^*/(++)	−/0
B: physics family vs agnostic	0/+	++/++	0/++	0/++	0/++	0^*/0	0/++
C: physics+coverage vs random	0/+	0/++	−/++	+/++	0/++	0^*/++	−/(+)
C: physics+coverage vs det. cov.	0/0	0/0	0/(+)	0/(+)	0/(+)	0^*/(−)	0/0
D: M1 Lagrangian	++/—	—/—	0/(−)	—/0	++/—	0^*/(−)	++/(−)
E: M2 primal–dual	0/0	0/0	+/—	−/(—)	0/(—)	0^*/(+)	0/(−)
F: DualWeight	0/0	0/0	0/0	0/(−)	0/0	0^*/(−)	0/0
F: LRA	0/−	—/—	0/—	—/—	0/—	0^*/—	0/(−)
F: GradNorm	0/0	0/0	0/0	0/0	0/0	0^*/0	0/0
F: SoftAdapt	0/0	0/(−)	0/(−)	0/(—)	0/(−)	0^*/(—)	0/0
F: ReLoBRaLo	0/0	0/0	0/(−)	0/(—)	0/(−)	0^*/(−)	0/0
G: Adam→L-BFGS-B (P+C set)	0/(−)	0/0	0/+	0/++	0/—	0^*/—	0/0
G: Adam→L-BFGS-B (O1 median)	0/−	0/+	0/++	0/++	0/—	0^*/—	0/−
H: $I1$ prior, A0, 10%	0/(+)	(+)/(+)	0/0	0/0	0/0	0^*/0	0/0
H: $I1$ prior, A0, 2%	0/0	0/0	0/(−)	(−)/(—)	0/(−)	0^*/(—)	0/0
H: $I1$ prior, plain64, 10%	(+)/(+)	(++)/(++)	0/(++)	0/(++)	0/(++)	(++)/(+)	0/0
I: no affine skip, 10%	(++)/(—)	(−)/(—)	(+)/(—)	(−)/(—)	(++)/(—)	$(—^*)$/(—)	(+)/(—)
I: medium256, 10%	(++)/(++)	(++)/(++)	(++)/(++)	(++)/0	(++)/(++)	$(—^*)$/(++)	(++)/0
I: medium256, 2%	0/(—)	(−)/(++)	0/(++)	(−)/(−)	0/(—)	0^*/(—)	0/(—)
I: width256, 10%	(+)/(++)	0/(++)	0/(+)	(−)/(+)	(+)/(++)	0^*/(++)	0/(+)
I: depth6, 10%	0/0	(—)/(—)	(—)/(—)	(−)/(—)	0/(—)	0^*/(—)	(−)/(—)
I: SiLU, 10%	0/(++)	0/(−)	0/(—)	(—)/(−)	0/(+)	0^*/(—)	0/0
I: ReLU, 10%	0/(—)	(—)/(—)	0/(—)	(—)/(—)	0/(—)	0^*/(—)	0/(—)
J: Experiment 4 optimal DoE	? (planned protocol; not evaluated in this manuscript)						

Table 11 Level 1 (global) physical effect of each training intervention: mean relative change (%) against the matched comparator, with the symbol rule of Table 10 (Holm family: the global quantities of that intervention). For stability preservation a positive change is an improvement, so its symbol is reported with the sign of the error convention reversed in the text. Trajectory error was not stored by Experiment 3 (?).

Training process	E_F	$E_{A,\mathrm{rel}}$	E_Φ mean	E_Φ median	$\lvert\Delta\alpha\rvert$	stab. pres.	E_{traj} long
A: det. coverage vs random	-5.3 (+)	+3.6 0	-19.9 ++	-15.8 ++	-35.4 ++	+16.5 (++)	-22.8 (++)
B: physics family vs agnostic	-14.3 ++	-2.1 0	-5.8 (+)	-4.3 0	-15.8 ++	+15.4 ++	-22.6 ++
C: physics+coverage vs random	+2.5 0	+3.2 0	-18.7 ++	-15.1 ++	-35.4 ++	+21.5 (++)	+18.7 (—)
C: physics+coverage vs det. cov.	+9.7 (−)	-0.4 0	+1.6 0	+1.5 0	+1.0 0	+4.1 0	+53.8 (—)
D: M1 Lagrangian	+9.9 (−)	-16.5 ++	+27.6 —	+22.4 —	+63.6 —	-17.2 —	-9.7 (+)
E: M2 primal–dual	+1.1 0	-2.6 0	+8.6 −	+11.3 —	+30.5 —	-10.9 (—)	+17.6 (—)
F: DualWeight	+5.0 (−)	+0.5 0	+3.3 0	+0.4 0	+5.4 (−)	+1.5 0	-7.7 (+)
F: LRA	-5.0 0	+0.6 0	+8.5 −	+11.5 —	+17.0 (—)	-9.6 (−)	-3.2 0
F: GradNorm	+1.1 0	+0.2 0	+0.1 0	-2.7 0	+2.1 0	+1.2 0	-0.5 0
F: SoftAdapt	+3.5 0	+0.9 0	+9.6 (−)	+7.3 (−)	+8.9 (−)	+3.5 0	-13.5 (++)
F: ReLoBRaLo	+5.6 (−)	+0.5 0	+4.6 0	+2.1 0	+2.8 0	+3.9 0	-4.0 0
G: Adam→L-BFGS-B (P+C set)	+21.6 —	+0.8 0	+3.0 0	+1.0 0	+33.0 (—)	-13.2 (—)	+1.4 0
G: Adam→L-BFGS-B (O1 median)	+18.9 —	+0.5 0	+1.3 0	-1.7 0	+16.0 —	-6.2 (−)	-5.8 (+)
H: $I1$ prior, A0, 10%	+1.6 0	+0.7 0	-7.9 (+)	-6.3 (+)	-52.5 (++)	+7.4 (+)	?
H: $I1$ prior, A0, 2%	+5.0 (−)	+0.4 0	+13.4 (—)	+4.1 0	-12.9 (++)	+7.1 (+)	?
H: $I1$ prior, plain64, 10%	-2.3 0	-3.3 0	-25.9 (++)	-24.8 (++)	-18.8 (++)	+2.1 0	?
I: no affine skip, 10%	-16.1 (++)	-10.2 (++)	+35.7 (—)	+33.9 (—)	-2.2 0	+1.3 0	?
I: medium256, 10%	-67.9 (++)	-57.7 (++)	+4313.0 (—)	-14.2 (++)	+565.9 (—)	-9.6 (−)	?
I: medium256, 2%	+7.6 (−)	+1.2 0	-2.3 0	-6.1 (+)	+20.1 (—)	+23.2 (++)	?
I: width256, 10%	-9.4 (+)	-4.4 0	-12.9 (++)	-16.4 (++)	+31.6 (—)	-0.5 0	?
I: depth6, 10%	+0.6 0	+3.5 0	+30.0 (—)	+27.1 (—)	+42.2 (—)	-2.8 0	?
I: SiLU, 10%	-5.9 (+)	-3.2 0	+3.8 0	+10.5 (—)	+17.1 (—)	+2.6 0	?
I: ReLU, 10%	+4.5 0	+2.3 0	+40.7 (—)	+36.0 (—)	+150.9 (—)	-6.7 (−)	?

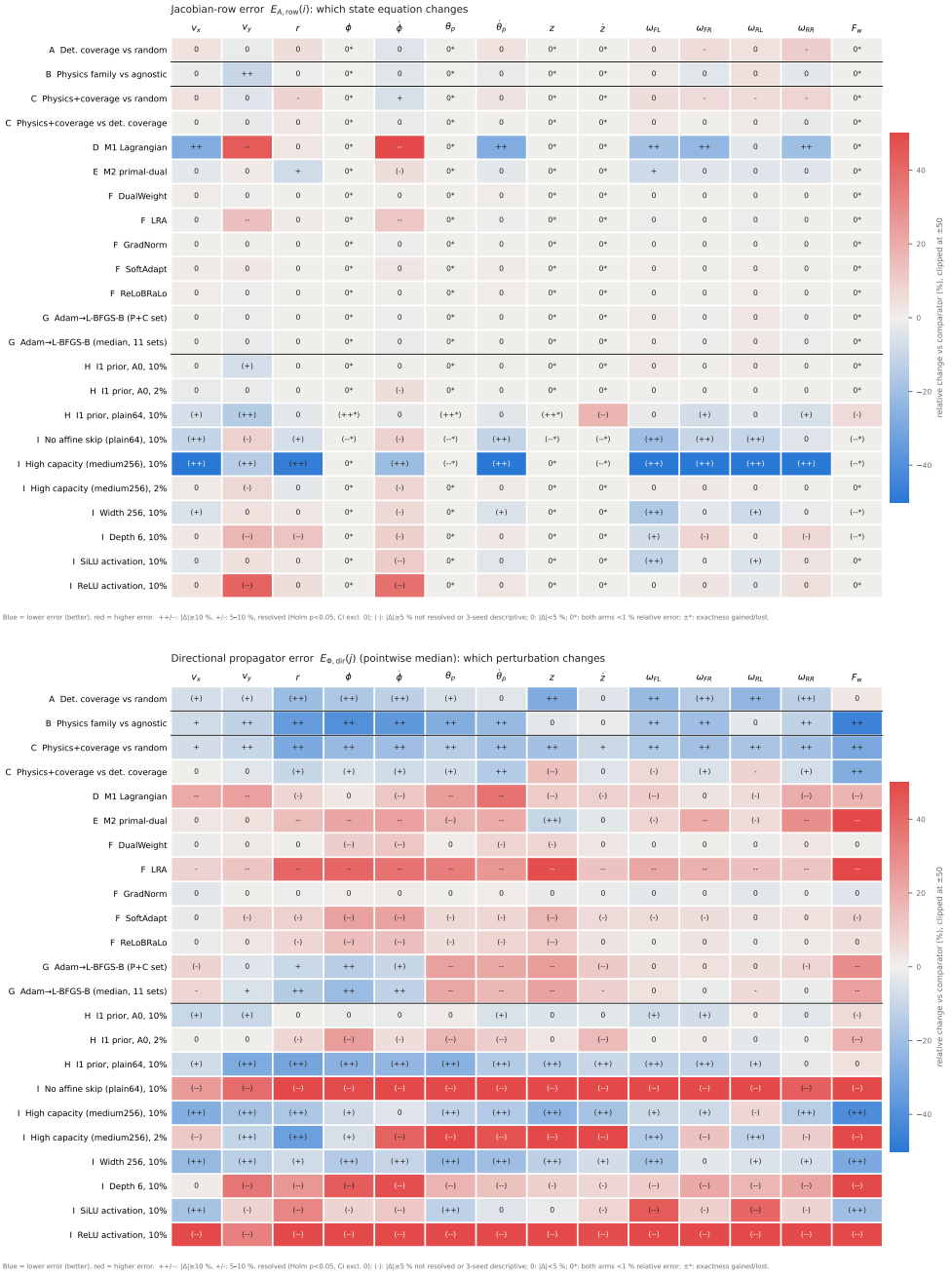


Fig. 1 Which training process changes which physical state. Rows: training interventions against their matched comparators (v6/v5: ten seeds, post-hoc exploratory; Experiment 3: three seeds, descriptive; horizontal rules separate the evidence blocks). Columns: the 14 states. Cell colour: mean relative change of the metric (blue lower error, red higher error, clipped at ±50%); cell text: symbol of Table 10. Top: Jacobian-row error $E_{A,\mathrm{row}}(i)$ (local differential equation of state i). Bottom: directional propagator error $E_{\Phi,\mathrm{dir}}(j)$ (pointwise median) for a perturbation of state j.

Table 12 Where the analytical prior acts (Experiment 3, random selection, three seeds, descriptive). Top: missing-prior fraction $m_{ij} = -\langle \widehat{A}_{ij} - \bar{A}_{ij}, P_{ij} \rangle / \langle P_{ij}, P_{ij} \rangle$ at the seven Jacobian entries supplied by F_{prior} (P: scaled prior Jacobian; $m = 1$: the learned model lacks the whole prior term, $m = 0$: correct, and under $I1$ $m > 0$ means the learned residual cancels that share of the supplied prior). Bottom: Jacobian-row error $E_{A,\text{row}}$ of the rows receiving the prior (v_x, v_y, ϕ, θ_p, z) and of rows that do not.

Arch., budget	prior	$\partial \dot{v}_x / \partial v_y$	$\partial \dot{v}_x / \partial r$	$\partial \dot{v}_y / \partial v_x$	$\partial \dot{v}_y / \partial r$	$\partial \dot{\phi} / \partial \dot{\phi}$	$\partial \dot{\theta}_p / \partial \dot{\theta}_p$	$\partial \dot{z} / \partial \dot{z}$
A0, 10%	I0	-0.01	+0.55	+0.55	+0.12	-0.00	-0.00	-0.00
A0, 10%	I1	+0.47	-0.03	+0.24	+0.01	+0.00	+0.00	-0.00
A0, 2%	I0	+0.77	+0.86	+1.32	+0.23	-0.00	+0.00	-0.00
A0, 2%	I1	+0.32	-0.10	+0.57	+0.07	+0.00	+0.00	+0.00
plain64, 10%	I0	+0.65	+0.41	+0.11	+0.12	+0.00	-0.01	+0.02
plain64, 10%	I1	-0.02	-0.07	+0.04	-0.01	+0.00	+0.00	-0.00
plain64, 2%	I0	+0.17	+0.70	+0.84	+0.25	+0.09	+0.04	+0.04
plain64, 2%	I1	-0.39	+0.18	+0.38	+0.07	-0.00	-0.00	-0.00
medium256, 10%	I0	+0.22	+0.29	+0.04	+0.05	-0.00	+0.00	+0.00
medium256, 10%	I1	+0.23	-0.15	-0.19	+0.01	-0.00	+0.00	-0.00

		v_x	v_y	ϕ	θ_p	z	r	$\dot{\phi}$
A0, 10%	I0	1.010	0.299	0.000	0.001	0.000	0.834	0.367
A0, 10%	I1	1.013	0.278	0.000	0.000	0.000	0.833	0.366
A0, 2%	I0	1.188	0.618	0.000	0.001	0.000	0.951	0.694
A0, 2%	I1	1.168	0.604	0.000	0.000	0.000	0.947	0.729
plain64, 10%	I0	0.898	0.325	0.228	0.314	0.268	0.779	0.399
plain64, 10%	I1	0.839	0.278	0.000	0.000	0.000	0.760	0.390
plain64, 2%	I0	0.817	0.774	0.317	0.340	0.314	0.894	0.876
plain64, 2%	I1	0.762	0.642	0.000	0.001	0.000	0.887	0.875
medium256, 10%	I0	0.429	0.257	0.005	0.005	0.004	0.441	0.290
medium256, 10%	I1	0.444	0.227	0.002	0.003	0.003	0.466	0.305

Table 13 Where the learned model is actually nonlinear (rows with state-dependent reference dynamics). $V_i = \left[\text{mean}_q \|\widehat{\widehat{A}}_{i,:}(q) - \overline{\widehat{\widehat{A}}}_{i,:}\|^2 / \text{mean}_q \|\widehat{\widehat{A}}_{i,:}(q)\|^2\right]^{1/2}$ is the state-dependence of learned row i ($V_i = 0$: the row is affine/constant; compare V_i^{ref}), and $r_{\text{NN},i} = \text{mean}_q \|A_{\text{NN}}[i,:]\| / \|\widehat{A}[i,:]\|$ is the neural share of the row. Entries: V_i / $r_{\text{NN},i}$. The rows $\phi, \theta_p, z, \dot{z}, F_{\text{w}}$ are constant in the reference ($V^{\text{ref}} = 0$) and are omitted. Experiment 3 rows: three seeds, random selection, $I0$; descriptive.

	v_x	v_y	r	$\dot{\phi}$	$\dot{\theta}_p$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}
reference V_i^{ref}	0.60	0.68	0.73	0.70	0.60	0.68	0.59	0.75	0.68
A0 tanh, 2%	0.05/0.11	0.45/0.80	0.10/0.12	0.49/0.78	0.05/0.09	0.06/0.10	0.04/0.07	0.07/0.12	0.04/0.06
A0 tanh, 10%	0.12/0.24	0.66/0.86	0.32/0.42	0.69/0.83	0.10/0.21	0.15/0.23	0.07/0.15	0.10/0.19	0.07/0.15
SiLU, 10%	0.15/0.25	0.63/0.86	0.33/0.38	0.67/0.82	0.15/0.24	0.24/0.34	0.10/0.16	0.12/0.21	0.05/0.11
ReLU, 10%	0.13/0.20	0.63/0.85	0.32/0.36	0.66/0.81	0.12/0.20	0.17/0.24	0.10/0.14	0.17/0.24	0.09/0.13
depth6, 10%	0.32/0.36	0.67/0.86	0.58/0.55	0.68/0.82	0.31/0.34	0.46/0.48	0.33/0.28	0.47/0.44	0.32/0.29
width256, 10%	0.23/0.33	0.67/0.85	0.49/0.53	0.70/0.83	0.23/0.31	0.36/0.47	0.17/0.22	0.30/0.37	0.12/0.18
medium256, 2%	0.08/0.10	0.45/0.69	0.16/0.18	0.57/0.72	0.08/0.09	0.10/0.11	0.09/0.09	0.12/0.15	0.08/0.10
medium256, 10%	0.51/0.72	0.67/0.87	0.68/0.77	0.69/0.89	0.51/0.70	0.62/1.84	0.53/1.01	0.71/2.19	0.58/1.19
plain64 (no skip), 10%	0.25/1.00	0.67/1.00	0.38/1.00	0.69/1.00	0.26/1.00	0.36/1.00	0.21/1.00	0.32/1.00	0.17/1.00
v6 A0, physics+cov., Adam	0.06/0.05	0.70/0.68	0.36/0.27	0.69/0.69	0.05/0.04	0.06/0.05	0.04/0.04	0.06/0.06	0.04/0.04
v6 A0, physics+cov., L-BFGS	0.16/0.15	0.73/0.68	0.55/0.40	0.72/0.70	0.16/0.14	0.22/0.19	0.14/0.12	0.25/0.21	0.20/0.17

6.11 Cross-experiment synthesis

The three experiments answer different questions and should not be collapsed into a single winner/loser narrative.

First, reference metrics can be diagnostically useful without implying that the same metrics are optimal for training. Experiment 1 shows a strongly budget-dependent validation effect: C_{top3} leads at 2% and 5%, while symmetric growth reaches the strongest observed macro recall at 10% (0.4338 versus 0.1011 for random sampling). Experiment 2, however, shows that the strongest 14-state Jacobian training results come from individual physical selectors rather than the combined training policy.

Second, observational bias changes learned physical fidelity, but the effect is metric-, model-, and policy-dependent. The primary family-level comparison shows a statistically reliable but practically small Jacobian gain, while the per-strategy decomposition shows that physics-plus-coverage is not the source of that gain. Its finite-time advantage is largely reproduced by deterministic coverage alone.

Third, architecture and analytical prior affect the same fidelity metrics, but the screening does not support a simple "larger network is better" or "more physics is better" rule. Most skip architectures remain close to an affine-Jacobian baseline; removing the affine skip and, separately, providing very high capacity with enough data are the two clearest structural changes. Informativeness is therefore not solely a property of a data point. Its value depends on the remaining point set, the learner parameterization, the target metric, and the available budget.

Fourth, interventions with similar global effects act on different physics. M1 and M2 share an $E_{A,\text{rel}}$-better/E_{Φ}-worse pattern but change different equations and couplings. Loss balancing can act only on the neural lateral/roll subspace. L-BFGS-B changes no equation but redistributes finite-time fidelity. The analytical prior and the architecture decide which equations the data can shape at all.

7 Discussion

This section interprets the empirical findings in relation to the broader scientific-machine-learning literature. The emphasis is on what the experiments establish about trajectory fidelity, observational bias, architecture, and learner adequacy, while keeping retrospective and screening evidence separate from the fresh-seed held-out results.

7.1 Trajectory accuracy and physical fidelity are distinct

The strongest confirmatory result in the paper is the trajectory-versus- physical-fidelity result reported in Sec. 6.2. Low trajectory error can coexist with high Jacobian or propagator error on fresh held-out operating points. This supports the use of a hierarchy of nominal, local, and finite-time metrics rather than treating trajectory RMSE as a complete validation criterion.

The result does not imply that trajectory accuracy is unimportant. Instead, trajectory error answers a different question: whether nominal predictions are accurate. Jacobian and frozen-local propagator errors ask whether the learned model reproduces

how nearby states separate or reorient. A model intended for sensitivity analysis or closed-loop simulation should be evaluated at both levels.

7.2 Statistical and practical interpretation of the observational-bias effect

The primary observational-bias result in Sec. 6.3 and Table 2 illustrates why practical thresholds should be separated from statistical significance. The Jacobian improvement is statistically convincing after Holm correction, but a 2.09% relative reduction is smaller than the predefined 5% practical effect. Conversely, the propagator reduction is practically larger (5.80%) but does not pass the corrected significance criterion. The correct journal-level interpretation is therefore not a binary success/failure statement: the experiment identifies a small but statistically stable Jacobian effect and a larger but less certain finite-time effect.

This split result is compatible with the active-learning literature, where sample informativeness, representativeness, and diversity are separate design criteria rather than a single scalar objective [7]. It is also consistent with importance-sampling results for physics-informed learning, which show that changing the collocation distribution can improve convergence without implying that every physical quantity improves uniformly [6].

7.3 Observational bias is mechanism- and model-dependent

The model-complexity results in Sec. 6.4 and Table 3 show that a physics-informed sampling strategy should not be treated as a model-independent intervention. The 3-state learner benefits more strongly, the 14-state learner improves modestly, and the 7-state Jacobian result reverses sign. The held-out manoeuvre analysis in Sec. 6.5 leads to the same mechanism-dependent interpretation: $E_{A,\mathrm{rel}}$ improves across all five manoeuvres, whereas E_{Φ} improves only for two.

This pattern is consistent with the many-to-many correspondence between reference metrics and validation errors. A score such as I_J, symmetric growth, or g_F emphasizes a specific local mechanism. The learned Jacobian, however, aggregates all state couplings, and the frozen-local propagator further aggregates those local effects through the matrix exponential. A sampling strategy can therefore improve one mechanism without improving the complete fidelity vector. This interpretation is aligned with nonlinear system-identification practice, where useful excitation must be judged jointly with model structure and operating-domain coverage [9], rather than by informativeness alone.

7.4 Different metrics answer different physical validation questions

The post-hoc decomposition in Sec. 6.10 clarifies why no single scalar metric should be treated as a complete physical validation criterion. Each metric weights a different dynamical mechanism and therefore changes the physical meaning of "good" validation. This distinction also separates the *reference metrics used to select informative operating*

points from the *error metrics used to validate a learned model*. A selector can be highly informative for one mechanism without producing a model that is uniformly better under all validation metrics.

The M1 result provides a concrete example of why this hierarchy matters. $E_{A,\mathrm{rel}}$ rewards the large improvement in wheel/pitch blocks because those blocks dominate the Frobenius discrepancy. Directional E_Φ, modal frequency/damping errors, and the spectral abscissa expose a different aspect: smaller lateral and roll coupling errors can be amplified through the matrix exponential and materially alter finite-time perturbation response. Thus a model can be *better under a global local metric* and simultaneously *worse under a physically important finite-time metric*.

The same principle changes how the reference selection metrics should be read. A high I_J score identifies rapid variation of the tangent map; symmetric growth identifies instantaneous amplification; g_F^C identifies transverse reorientation; non-normality identifies susceptibility to transient growth; finite-time amplification identifies strong response over a specified horizon; and the tyre-transition score identifies a particular nonlinear vehicle mechanism. None of these quantities is a generic "importance" measure. Their value depends on the downstream validation target. This explains why the best selector can change with validation budget in Experiment 1 and why mechanism-specific selectors can outperform combined scores for one training metric while losing for another.

Accordingly, validation should be reported as a vector of physically interpretable outcomes rather than collapsed prematurely to one mean:

$$\boxed{\mathcal{V}_{\mathrm{phys}} = [E_{\mathrm{traj}}, E_F, E_{A,\mathrm{rel}}, E_{\mathrm{block}}, E_{\mathrm{mode}}, E_{\Phi,\mathrm{dir}}, E_\Phi, E_{\mathrm{tail}}]\,.} \tag{84}$$

The exact subset should be matched to the intended use of the learned model. For nominal simulation, trajectory and vector-field accuracy may dominate. For sensitivity analysis or local control design, Jacobian blocks and modal quantities are more directly relevant. For perturbation propagation, closed-loop robustness studies, or safety-oriented simulation, directional and finite-time metrics become necessary. This target-dependent interpretation is more informative than declaring one metric universally primary.

7.5 How training interventions redistribute physical fidelity

The state-resolved analysis in Sec. 6.10 shows that training interventions redistribute fidelity across state equations, couplings, modes, and perturbation directions rather than simply moving one scalar error. The mechanism depends on what part of training changes. Selection changes which regimes determine the fitted affine slopes and neural residual; loss reweighting changes the residual updates on a fixed point set; optimizer refinement changes the solution reached for the same objective; and architecture or analytical priors change which tangent structures can be represented efficiently.

M1 gives the clearest example. Its 16.5% reduction in $E_{A,\mathrm{rel}}$ is obtained mainly from large longitudinal, pitch-related, and wheel blocks, while the lateral-velocity and roll-rate equations, their influence on other states, and lateral/yaw modes deteriorate.

Thus a model preferred by global Frobenius Jacobian error may be useful for wheel-slip sensitivity yet unsuitable for lateral-stability or roll-response studies. Conversely, physics-family selection changes global $E_{A,\mathrm{rel}}$ little but improves several body-attitude perturbation directions, whereas Adam→L-BFGS-B leaves the state equations nearly unchanged while redistributing finite-time fidelity and introducing spurious couplings.

This motivates reporting state-resolved quantities together with the global metrics. A Jacobian-row error asks whether the differential equation of one state is correct; a column error asks whether that state influences the rest of the vehicle correctly; and a directional propagator error asks whether a perturbation of that state is propagated correctly over the frozen-local horizon. The global $E_{A,\mathrm{rel}}$ and E_Φ remain useful summaries, but they should not replace these task-relevant state and coupling diagnostics.

7.6 Why more physics-selected points do not necessarily produce a better learner

The combined evidence from the per-strategy training results in Sec. 6.6, the adaptive-allocation extension in Sec. 6.7, and the architecture/bias screening in Secs. 6.8–6.9 argues against interpreting the physics score, by itself, as a direct measure of training value. The v6 extension adds an important qualification: the earlier failure of physics-heavy fixed mixtures is not evidence that a physics-heavy set must fail. Rather, it shows that a fixed fraction is a poor proxy for the joint requirements of informativeness and coverage. The clearest indication comes from the development sweep used to freeze the observational treatment in Experiment 3. As the fraction of the budget assigned to the physics-composite selector increased from 25% to 100%, the multi-metric development criterion became progressively worse relative to random sampling, changing from -3.2% at $\alpha = 0.25$ to -20.7% at $\alpha = 1$. The retained 25% physics and 75% coverage mixture is therefore best interpreted as the least harmful candidate among the tested mixtures, not as evidence that a particular fraction of physics-selected points is intrinsically optimal.

Experiment 2 reaches the same conclusion through the held-out per-strategy results in Sec. 6.6 and Table 4. In the 14-state, 5% condition, `physics_plus_coverage` does not improve E_A relative to the physics-only composite, is worse than random sampling, and differs only marginally from deterministic coverage. Individual mechanism-specific selectors, especially tyre transition, non-normality, and finite-time amplification, produce lower Jacobian errors in the exploratory ranking. The finite-time result is different: at 5% and 10% the mixed selector improves E_Φ relative to random and the physics-only composite, yet it is statistically indistinguishable from deterministic coverage. This indicates that its propagator advantage is explained primarily by the broad coverage component rather than by an independent benefit from adding high-score physics points.

These findings clarify why valid, low-score points should not be described as "bad" data. A point located in a nominal or weakly nonlinear region can still anchor the learned vector field, connect separated high-curvature regions, and reduce extrapolation over the admissible domain. Conversely, a highly ranked point can consume part of a limited budget without improving the particular fidelity quantity used for evaluation. Training-set quality is therefore a property of the collection of points and its interaction

with the learner, rather than a property of each point considered in isolation. In the present experiments, the useful question is not how many high-score physics points can be inserted, but how the available budget should be allocated between mechanism-specific informativeness and representative coverage for a given model dimension, architecture, and validation target.

This distinction also prevents a misleading interpretation of the architecture screening. A positive observational–inductive interaction does not imply that the combined condition is automatically the best model; mathematically, a positive interaction can arise even when both individual interventions are harmful and the joint condition is merely less harmful than the additive expectation. The experiments therefore support a conditional, target-dependent training policy rather than a monotonic "more physics is better" rule.

7.7 Coverage should be enforced as a constraint, not inferred from a sampling fraction

The adaptive-selection results in Sec. 6.7 refine the paper's earlier interpretation of the physics/coverage trade-off. The frozen v5 policy represents coverage indirectly by reserving 75% of the sample budget for deterministic coverage. Its development sweep showed that simply increasing the physics fraction under that union rule made the multi-metric criterion progressively worse. M1 changes the mathematical problem: instead of fixing how many points must come from each source, it maximizes the physics score while requiring the selected set itself to retain at least 90% of the facility-location coverage achieved by deterministic coverage. The complete selection procedure is given step by step in Sec. 5.3.1. In the primary cell, this construction permits a 99.4% physics-labelled set and still lowers Jacobian error by 16.5% relative to the frozen-alpha comparator.

This result shows that the relevant quantity is not the nominal proportion of physics-labelled samples but the geometric coverage of the selected set. A physics-ranked point can simultaneously contribute to the coverage objective; conversely, reserving a large fraction of the budget for nominal coverage does not guarantee that the resulting union is optimal for the target derivative metric. The appropriate conceptual separation is therefore

$$\boxed{\begin{array}{c}\text{physics fraction} \neq \text{physical informativeness},\\ \text{physical informativeness} \neq \text{operating-domain coverage}.\end{array}} \tag{85}$$

The first quantity is an allocation parameter, whereas the latter two are properties of the resulting selected set.

The strong M1 Jacobian result does not establish a universally superior training strategy because the same selector increases E_Φ by 27.6%. This trade-off is scientifically important: a set can be well covered in the geometric facility-location sense and highly informative for recovering local Jacobian structure while still yielding poorer finite-time propagation. The coverage constraint in Eq. (45) controls where samples lie, not whether all dynamical mechanisms relevant to the matrix exponential are learned correctly. Selection should therefore be posed as a target-specific constrained

design problem, and the held-out validation vector must contain every fidelity quantity required by the downstream application.

The negative loss-balancing and optimizer results reinforce the same point. Once the point set is fixed, changing the relative training weight of PHY and COV samples does not reproduce the benefit of M1, and adding L-BFGS-B after Adam does not improve the primary structural metrics. At state level the balancers act only on the neural lateral/roll subspace and improve no subsystem, whereas L-BFGS-B leaves every state equation unchanged while introducing spurious couplings and trading yaw/roll against pitch/heave finite-time fidelity (Sec. 6.10). For the present learner, selection geometry has a substantially larger observed effect than these post-selection optimization modifications. This finding supports solving the data-allocation problem before adding optimizer complexity.

7.8 Architecture is part of the physical-learning problem

Experiment 3 makes explicit that architecture is not merely a nuisance variable, but the effect is more specific than a generic depth/width ranking. The screened learner contains a fitted affine output model plus a neural remainder, except for `plain64`. Because most networks start with a small neural output and are trained only on vector-field residuals, many fits remain close to the local derivative structure of the affine component. This explains why most architectures cluster near mean $E_A \approx 0.92$–1.01 even when their final vector-field losses differ substantially.

Two cases depart from this pattern. `plain64` removes the affine skip, forcing the neural network to represent the full field and its local slopes; it reaches substantially lower mean Jacobian error across the balanced cells. `medium256` becomes much better only at the 10% budget, suggesting that very large capacity is useful for local Jacobian learning only when enough training data are available. Activation changes, additional depth, and moderate width alone show much smaller structural effects in this screening.

The state-resolved decomposition sharpens this interpretation. With the affine skip and vector-field loss, the network uses its nonlinear capacity almost exclusively in the lateral and roll-rate equations, while the longitudinal, pitch-rate, and wheel equations stay nearly affine (Table 13). High capacity combined with enough data is the only screened change that makes those stiff equations state dependent, and it improves body and wheel equations together. Removing the skip makes every row neural but loses the exact linear kinematics and degrades every perturbation direction.

The analytical prior $I1$ is likewise targeted rather than universal. With an affine skip its kinematic integrators are redundant, because the affine map already reproduces those linear rows. The rotating-frame bilinear terms it supplies are entries that the $I0$ learner largely misses, but the learned residual partly cancels them, and tyre terms dominate the v_x and v_y rows. The global E_A effect is therefore small. Without the affine skip the prior also restores the exact kinematics and improves lateral, yaw, and roll propagation by 26–31%. The more frequent median-E_Φ benefit is therefore an architecture-conditioned effect of specific exact couplings, not a general consequence of adding physics.

These results are compatible with prior work showing that activation and parameterization can strongly condition scientific neural-network training [4, 5]. They do

not, however, justify a general claim that deeper, wider, affine-skip, or physics-biased architectures are superior. The supported screening interpretation is conditional: architecture controls which local structures the learner can absorb efficiently from a given point set.

7.9 Finite-time fidelity requires robust summaries

The architecture screening reveals an important measurement issue for finite-time fidelity. Because $\Phi = \exp(\bar{A}T)$, a single spurious Jacobian eigenvalue with large positive real part can make E_Φ enormous. Several large tanh models produce isolated values between approximately 10^5 and 10^{11}, while their typical per-point propagator errors remain close to one. A mean over points therefore conflates typical fidelity with rare local instability.

For architecture comparisons we accordingly report median-based E_Φ as the typical-fidelity statistic and preserve Q99 or maximum error as a separate stability outcome. This is not a cosmetic change: at 2%, the number of architectures for which the inductive prior improves E_Φ changes from 7 under the mean to 3 under the median, while the favourable interaction count changes from 6 to 10. Conclusions about finite-time fidelity must therefore state the aggregation rule explicitly.

7.10 Training-point selection and validation-point selection are not interchangeable

Experiment 1 and Experiment 2 address different causal questions, as shown by the validation-selection results in Sec. 6.1 and the training-selection results in Secs. 6.3 and 6.6. A validation selector is evaluated with a fixed learned error field:

$$\mathcal{S}_{\text{val}} \longrightarrow \text{critical-region coverage.} \tag{86}$$

A training selector changes the model itself:

$$\mathcal{S}_{\text{train}} \longrightarrow \theta^\star \longrightarrow E_k(\boldsymbol{q}; \theta^\star). \tag{87}$$

The fact that C_{top3} has the highest 5% validation recall in Table 1 therefore does not imply that the same selection rule must train the best model. This is especially important for C_{top3}: it is a frozen retrospective validation selector and is distinct from the later six-metric `physics_composite`. Its validation recall cannot be used as evidence that a three-metric training mixture is optimal. The architecture screening further shows that the effect of a training distribution depends on the learner receiving it.

7.11 Best-supported training protocol from the present experiments

The experiments do not identify a universally best neural network, but they do support a defensible procedure for designing and selecting a learned vehicle-dynamics model.

The first decision should be the physical fidelity quantity that the application actually requires. If the intended use is primarily nominal prediction, held-out vector-field and trajectory errors are the most direct selection criteria. If the learned model will also be used for sensitivity analysis, controller evaluation, local stability reasoning, or perturbation propagation, then the model-selection criterion must additionally include local Jacobian and finite-time errors. This ordering matters because the experiments show that a model can achieve a favorable trajectory error while still reproducing the local tangent structure poorly. Consequently, training-point selection should be matched to the fidelity objective rather than based on a generic physics score.

For Jacobian-oriented learning in the 14-state model, the exploratory per-strategy results in Sec. 6.6 identify `tire_transition`, `non_normality`, and `finite_time` as strong single-score baselines. The subsequent v6 result in Sec. 6.7 provides a stronger design principle: rather than fixing a physics fraction, maximize the normalized physics score subject to an explicit coverage constraint. In the primary 14-state, 5% cell, M1 reduces E_A by 16.5% relative to the frozen-alpha comparator while maintaining $\widehat{C}(S) \geq 0.90$. This makes coverage-constrained selection the strongest tested candidate for a Jacobian-centric training policy in the primary cell. It is not a universal winner, however, because its E_Φ error simultaneously increases by 27.6%. The choice of selector must therefore remain conditional on the physical-fidelity quantity required by the application. In contrast, `physics_plus_coverage` gives $E_A = 1.339$ at 5%, essentially matching `physics_composite` (1.338), exceeding the error of random sampling, and differing only slightly from deterministic coverage. The evidence therefore does not justify using the mixed selector as a default when the primary objective is Jacobian fidelity.

When the principal objective is finite-time fidelity, the same per-strategy results in Sec. 6.6 support a different emphasis. At 5% and 10%, `physics_plus_coverage` improves E_Φ relative to the physics-only composite and random sampling, but it is statistically indistinguishable from deterministic coverage. The practical implication is that broad trajectory and operating-domain coverage should be preserved before additional physics-ranked points are assumed to add value. Because a small number of locally unstable predictions can dominate the mean of E_Φ, the typical propagator error is more appropriately summarized by the pointwise median, while Q99 or the maximum should be retained as a separate indicator of rare local instability. This distinction between typical fidelity and instability risk is especially important for the larger tanh architectures, for which isolated propagator blow-ups are observed.

The architecture should then be selected using exactly the same point set so that data and model effects are not confounded, following the matched design described in Sec. 5.5 and interpreted through the screening results in Sec. 6.8. The screening results suggest that the most consequential architectural differences are not a smooth activation or depth ranking. Removing the fitted affine output skip in `plain64` forces the neural component to represent the entire local slope field and yields substantially better Jacobian fidelity than most skip architectures in the screening. A second distinct regime appears for `medium256`, which becomes markedly better for E_A at the 10% budget but not at 2%, indicating that additional capacity becomes useful only when the training set is large enough to constrain it. By comparison, activation substitutions

and moderate depth or width changes produce smaller differences. The present evidence therefore supports comparing a no-skip baseline and a sufficiently expressive model under matched data, rather than declaring tanh, SiLU, GELU, ReLU, or softplus universally preferable.

When exact analytical terms are known and available at inference time, the $I1$ formulation defined in Eqs. (33)–(35) and tested under the matched design of Sec. 5.5 provides a conservative way to introduce physics without changing the supervised objective. The learned model is written as

$$\widehat{\boldsymbol{F}}(\boldsymbol{x}, \boldsymbol{u}) = \boldsymbol{F}_{\text{prior}}(\boldsymbol{x}) + \widehat{\boldsymbol{R}}_\theta(\boldsymbol{x}, \boldsymbol{u}), \tag{88}$$

so that the network represents only the residual dynamics after subtracting the rotating-frame bilinear terms and the selected kinematic integrators. Because this prior is additive and the residual can in principle cancel it, it acts as a targeted inductive bias rather than a hard constraint. Its effect on global Jacobian error is small in the screening, although typical propagator fidelity improves more often at the 10% budget. The appropriate interpretation is therefore that an exact prior can simplify part of the representation, but it cannot compensate for inadequate coverage, insufficient model capacity, or a poorly chosen training-point policy.

All completed training experiments optimize the same normalized vector-field mean-squared error,

$$\mathcal{L}_F = \frac{1}{Nn} \sum_{i=1}^{N} \left\| \left[\boldsymbol{F}(\boldsymbol{q}_i) - \widehat{\boldsymbol{F}}_\theta(\boldsymbol{q}_i) \right] \oslash \boldsymbol{\sigma}_F \right\|_2^2 . \tag{89}$$

Here $\mathcal{L}_F$ is the scalar training objective, N is the number of selected training points, n is the number of state derivatives, $\boldsymbol{q}_i$ is the i-th training operating point, $\boldsymbol{F}(\boldsymbol{q}_i)$ is its analytical reference target, $\widehat{\boldsymbol{F}}_\theta(\boldsymbol{q}_i)$ is the neural prediction, and $\boldsymbol{\sigma}_F$ is the componentwise target scale defined above. The operator $\oslash$ denotes componentwise division. This fact places an important limit on the recommendations that can be drawn from the study. Jacobian, directional-growth, g_F, and propagator losses were not used during training, and the proposed $I2$–$I5$ variants were not implemented. The current results therefore cannot establish that direct derivative or propagator supervision would improve the final model. Such losses are a logical next experiment, particularly because the present vector-field objective often leaves the learned Jacobian close to that of the fitted affine component, but they should remain clearly separated from the claims supported by the completed runs.

Taken together, the strongest training procedure supported by the present evidence is sequential rather than formulaic. Candidate data should first be restricted to the physically admissible domain and normalized consistently. Development and final test data should then be separated before any selector, mixture coefficient, architecture, or prior is chosen. On development data, random sampling and deterministic coverage should remain explicit baselines, and mechanism-specific selectors should be evaluated individually against the fidelity quantity that matters for the intended application. The selected point set should then be held fixed while candidate architectures or analytical

priors are compared under identical optimization settings. Only models that pass the learner-quality gate should proceed to detailed structural analysis, which should include at least vector-field error, Jacobian error, typical finite-time error, and a tail or maximum finite-time stability indicator. After these choices are frozen, the final evaluation should be performed once on the untouched test set.

This procedure also clarifies how the available training budget should be allocated. Invalid or inconsistent points should be removed before ranking, whereas valid low-score points should be retained whenever they improve domain support. A fixed physics/coverage coefficient such as the v5 α should be treated as a development hyperparameter rather than as a physical constant. More importantly, the v6 result suggests replacing a fixed fraction with an explicit coverage constraint whenever the target is Jacobian fidelity: the physics score can then be increased only while the selected set remains geometrically representative. If finite-time fidelity is the target, the selector must be re-evaluated because the M1 Jacobian gain is accompanied by a large propagator penalty. The current adaptive loss-balancing variants do not improve either target consistently, and the Adam-to-L-BFGS refinement provides no primary-cell benefit, so Adam remains the simpler supported optimization baseline for this study. In this sense, the best-supported strategy is target-matched, coverage-aware, and architecture-verified rather than based on maximizing the number or magnitude of physics-selected points.

7.12 Interpretation of the quality-gate failures

The v5 amendment described in Sec. 5.4 prevents strategy-induced failures from being hidden by complete-case filtering. The 21 failed primary-budget physics fits are retained as evidence that some point-selection strategies can make a learner structurally inadequate even when the same base learner succeeds on the agnostic data. This is scientifically informative: the quality gate is not only a preprocessing filter; it is itself an outcome. The need for such a gate is also supported by the PINN optimization literature: low composite loss can mask imbalanced convergence or poorly learned physical components [3, 4]. Our gate operationalizes the same concern for learned vehicle dynamics by requiring direct held-out structural checks.

The corrected supplementary interpretation is also important. A failed fit in one arm does not automatically make every supplementary cell descriptive-only. Under the v5 rule, the 7-state 1% cell remains inference-eligible because the failures occur only in the physics arm, whereas the 3- and 14-state 1% cells remain descriptive.

8 Limitations

This section delineates the scope of the claims. The experimental results are strongest for relative comparisons within the declared analytical vehicle models and weaker for statements that would require external simulator, real-vehicle, or fully confirmatory architecture evidence.

The reference throughout is an analytical hierarchy of 3-, 7-, and 14-state vehicle models. The results therefore establish fidelity relative to the implemented analytical

dynamics, not relative to CarMaker, CarSim, proving ground measurements, or real-vehicle data.

Experiment 1 is retrospective because its held-out results had already been inspected. Experiment 2 uses fresh operating points and fresh neural seeds, but the manoeuvre families were known from previous work; it is therefore a fresh-seed held-out replication rather than a completely untouched benchmark. Experiment 3 is explicitly development-stage screening with only three seeds and supports no confirmatory architecture claims.

The finite-time quantities reported here are frozen-local $\exp(\bar{\boldsymbol{A}}T)$ metrics at $T = 0.5$ s. They should not be interpreted as trajectory-varying tangent propagation. A future study could add time-ordered state-transition matrices along held-out trajectories.

The physics metrics and error norms depend on the declared state coordinates, normalization, admissible operating domain, manoeuvre library, and finite-time horizon. The conclusions are therefore conditional on these definitions.

The Experiment 3 inductive intervention is deliberately narrow: $I1$ adds only exact rotating-frame and kinematic terms and leaves all tyre-force and other nonlinear couplings to the learned residual. The study therefore does not test stronger physics supervision such as Jacobian, directional-growth, or propagator losses. In addition, the architecture study has only three seeds, so its contrasts remain screening evidence.

The v6 extension introduces two additional limitations. First, the M1 facility-location functional is a geometric coverage surrogate in normalized operating coordinates; satisfying $\widehat{C}(S) \geq 0.90$ does not guarantee coverage of every dynamical mechanism relevant to E_Φ, as demonstrated by the observed Jacobian/propagator trade-off. Second, the M2 implementation reuses the previous constraint residual when a sequential batch contains no physics-labelled points. A positive reused residual can drive the multiplier to its upper clip, so the primary M2 allocation collapse is not interpreted as a clean test of the ideal primal–dual method. The adaptive-method equations were documented after the run because the analysis-file hash was frozen in the experiment manifest; the numerical outputs themselves come from the saved fit records.

Two further analysis caveats should be stated explicitly. First, `finite_time` and `FTLE` select identical points because one score is a monotone transform of the other. The frozen Experiment 2 family analysis therefore contains 12 physics labels but 11 distinct selections and weights this selection twice in the family median. The primary family statistic is retained exactly as frozen; no post-hoc de-duplicated inferential estimate is substituted for that estimand in this manuscript. Second, the per-strategy rankings, affine-only comparisons, and physics-plus-coverage decomposition are post-hoc exploratory analyses of saved fits; they refine mechanism interpretation but do not replace the frozen confirmatory family-level test.

The state-resolved decompositions (rows, columns, blocks, directions, entries, modes, and the prior/affine/neural split) are post-hoc analyses of frozen checkpoints. For Experiments 2 and 2b they cover one cell (14 states, 5% budget, ten seeds) and are exploratory. For Experiment 3 they use three seeds and the development audit points and are descriptive. They do not create new confirmatory hypothesis families. With

64 blocks in a Holm family, no individual block can be resolved with ten seeds, so block results are descriptive. Mode labels depend on the declared scaling and dominant eigenvector component. The stability classification is tolerance sensitive because of near-zero reference modes. Single-block propagator injection is non-additive for stiff wheel/pitch dynamics. Mean pointwise row ratios are heavy-tailed for the wheel rows, so medians and aggregate ratios are also reported. The symbol rule was fixed before any contrast was computed, except for the treatment of relative errors below 1%, which was added after the first smoke test and before any text was written. Trajectory errors were stored only in aggregate, so no per-state trajectory attribution is made. The Experiment 2 family state profile mixes eleven distinct selections and is not attributed to any single selector. Finally, the M2 state signature reflects the implemented selector including its reused-constraint limitation. Experiment 4 is a planned protocol, and no aggregate or state-level Experiment 4 result is claimed.

9 Conclusion

The experiments show that physical fidelity in learned vehicle dynamics cannot be reduced to trajectory accuracy, to one global structural error, or to a single notion of "physics-informed" training. The fresh-seed training experiment provides the strongest direct evidence: low trajectory error can coexist with both high Jacobian error and high frozen-local propagator error. Nominal prediction, local tangent fidelity, and finite-time perturbation fidelity are therefore complementary validation levels, and the state-resolved analysis adds a fourth dimension: which physical state or coupling each level refers to.

What is observed. Training-point selection changes which physical regimes constrain the model. Reference-informed validation selectors concentrate the failures of an already trained model (macro recall 0.4338 for symmetric growth at 10%, against 0.1011 for random sampling). This is a validation property and does not transfer automatically to training. Physics-inspired training selection improves 14-state Jacobian error by only 2.09% in the confirmatory family test; in the post-hoc state-resolved decomposition, the corresponding gain is concentrated in the lateral-velocity equation. Its clearer, exploratory benefit is finite-time: perturbations of yaw, roll, and pitch propagate 22–40% more accurately. The coverage-constrained M1 selector lowers global Jacobian error by 16.5% by improving the longitudinal, pitch-rate, and wheel equations, which dominate the Frobenius norm. It degrades the lateral-velocity and roll-rate equations, their influence on the rest of the vehicle, and the lateral- and yaw-dominated modes, and so increases propagator error by 27.6%. M2 shows the same aggregate trade-off through a different state mechanism, and the finite-time advantage of the frozen physics-plus-coverage mixture is a coverage effect.

How it is learned. Loss weighting, optimizer, prior, and architecture determine how those observations are converted into a tangent map. With an affine output skip, data shape the stiff subsystems mainly through average slopes and the lateral/roll subsystem through the neural residual. Adaptive loss balancing changes that residual most visibly in the lateral/roll subspace and improved no subsystem. Adam→L-BFGS-B left every state equation unchanged but introduced spurious sensitivity to attitude,

heave, and disturbance states, and traded yaw/roll against pitch/heave finite-time fidelity. The exact analytical prior supplies rotating-frame couplings that a vector-field learner does not discover, but the learned residual partly cancels them. Its kinematic part is redundant when an affine skip is present, and its finite-time benefit is large mainly without that skip. Only high capacity combined with sufficient data made the longitudinal, yaw, pitch-rate, and wheel equations genuinely state dependent.

What is physically preserved. The answer differs by state, coupling, and time scale. A single intervention can improve the wheel and pitch-rate equations while degrading lateral and roll dynamics, leave every equation unchanged while changing which perturbations are propagated correctly, or improve local equations while introducing rare spurious unstable modes. The architecture and state-level results remain descriptive or exploratory. Within the declared analytical models they are nevertheless sufficient to show that a lower global metric is not evidence that the physics relevant to a given task has improved.

The best-supported procedure is therefore target-matched, coverage-aware, and architecture-verified. State the fidelity quantity and the physical states the application relies on. Retain random and deterministic-coverage baselines. Formulate physics-informed selection as a constrained design problem. Hold the point set fixed while architecture and priors are compared. Accept a model only after its held-out vector-field, Jacobian, finite-time, and state-resolved errors for those states are adequate. The resulting principle is

physical fidelity depends jointly on what is observed, how it is learned, which physical state or coupling is considered, and what is validated.

(90)

Physical prior knowledge should consequently be treated as a set of distinct design mechanisms—observation selection, loss weighting, optimization, analytical parameterization, architecture, and validation criteria—whose effects on each physical state and coupling must be measured rather than assumed.

10 Reproducibility and Provenance

The physical-interpretation analysis in Sec. 6.10 is a read-only post-processing analysis of the frozen v6 checkpoints and held-out data. All 110 checkpoint and parameter hashes in the analyzed primary cell were unchanged before and after the analysis, no training or selection entry point was called, and recomputed $E_{A,\mathrm{rel}}$ and E_Φ agree with the saved held-out errors to within 9.1×10^{-13}.

The state-resolved analysis of all training interventions (Sec. 6.10) extends this read-only layer to 716 frozen checkpoints (v5: 160, v6: 220, Experiment 3: 336). For every fit, the saved $E_{A,\mathrm{rel}}$, E_Φ, and E_F were reproduced to within 9.1×10^{-13}. The prior/affine/neural Jacobian decomposition sums to the protocol Jacobian to 4×10^{-16}. Checkpoint and parameter hashes were unchanged, no training or selection entry point was called, and the file trees of the three frozen experiments (30,703 files) were identical

before and after the analysis. The analysis configuration (groups, declared propagator entries, symbol rule) was written before any state-level result was computed and is archived with the code and all per-fit caches, contrast tables, and figures.

Experiment 1 reproduced its archived summary byte-for-byte. Experiment 3 reproduced the archived per-fit metrics. The Experiment 2 v5 analysis preserves the previous fits and thresholds while applying the disclosed gate interpretation described in Sec. 5. The v6 adaptive-selection and optimizer extension adds 2160 fits; its 360 Adam refits of the frozen v5 strategies are bit-identical to the corresponding v5 runs. The v6 audit reports no agnostic learner-gate failures by optimizer and retains 39 strategy-induced failures as evidence. The complete project report records the experiment directories, run manifests, audit outputs, compiled figures, and test results. Experiment 3 additionally stores all 336 seed-level fits, cell means, mean- and median-based factorial contrasts, and architecture summaries; Experiment 2 stores the complete per-strategy held-out table and the post-hoc physics-plus-coverage comparisons.

Declarations

Funding

This research was supported by the Software/Hardware Integration Lab (LISHA) at the Federal University of Santa Catarina (UFSC). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES).

Competing Interests

The authors declare that they have no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Ethics Approval and Consent to Participate

Not applicable. This study is based exclusively on analytical vehicle models, numerical simulations, and computational experiments. It does not involve human participants, human data, human samples, or live animals.

Consent for Publication

Not applicable. The manuscript does not contain identifiable personal data, images, or other material requiring consent for publication.

Data Availability

The research data generated and analyzed during this study consist of simulation outputs, trained-model checkpoints, evaluation results, and derived physical-fidelity metrics produced from the analytical vehicle-dynamics models and computational experiments described in the manuscript. The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Materials Availability

Not applicable. No physical materials, biological samples, or experimental specimens were used in this study. The work is based on analytical vehicle models, numerical simulations, and computational experiments.

Code Availability

The source code used to implement the analytical vehicle models, training and selection procedures, physical-fidelity metrics, computational experiments, and post-processing analyses is available from the corresponding author upon reasonable request.

Author Contributions

J.R.L.F. and A.A.F. contributed equally to the conception and design of the study, methodology development, experimental design, analysis and interpretation of the results, and preparation of the manuscript. Both authors contributed to the review and revision of the manuscript, approved the final version, and agree to be accountable for the content of the work.

Table 14 Physical interpretation of the main selection and validation metrics. No metric is treated as a complete description of physical fidelity.

Metric	Main physical question	Important limitation for interpretation
Trajectory / rollout error	Does the learned model reproduce the nominal state evolution over the tested trajectory?	Can remain small even when local sensitivities, modal structure, or perturbation growth are incorrect.
E_F, $E_{F,i}$	Nominal derivative fidelity: does the learned vector field reproduce the instantaneous derivative (of each state) at the operating point?	Measures nominal local prediction, not how the derivative changes with state or how perturbations evolve.
$E_{A,\mathrm{rel}}$	Does the learned model reproduce the complete local tangent map in the declared scaled coordinates?	The Frobenius norm is weighted toward large-magnitude blocks. In the present 14-state model it is dominated by wheel/pitch structure and can hide smaller lateral/roll couplings that are important for finite-time response.
Block/coupling error	Which physical subsystem or state-to-state coupling is learned incorrectly?	Requires a physically justified grouping and can be coordinate dependent; relative errors are ill-conditioned for structurally near-zero entries.
Jacobian-row error $E_{A,\mathrm{row}}(i)$	Is the local differential equation of state i correct (e.g. row v_y: lateral-acceleration dynamics; row $\dot{\phi}$: roll-acceleration dynamics; wheel rows: wheel-spin dynamics)?	Rows whose reference is linear are reproduced exactly by the affine skip and carry no information about training; mean pointwise ratios are heavy-tailed for wheel rows, so medians are also reported.
Jacobian-column error $E_{A,\mathrm{col}}(j)$	Is the influence of state j on the rest of the vehicle correct?	Detects spurious couplings that row errors miss; columns with small reference norm magnify absolute errors.
Row state dependence V_i, neural share $r_{\mathrm{NN},i}$	Is the learned equation of state i genuinely nonlinear, and is the nonlinearity carried by the network?	Describes the representation, not its accuracy; must be read together with the row error and the reference V_i^{ref}.
Spectral abscissa	Does the learned local linearization preserve the strongest exponential growth/decay tendency?	A thresholded stable/unstable label is sensitive to near-zero modes; the continuous abscissa error is more robust.
Matched ω_n, ζ	Are oscillatory local modes represented with the correct frequency and damping?	Eigenvalues must be matched, and mode labels based on dominant coordinates are basis dependent for mixed modes.
E_Φ	Does the learned local linearization reproduce total perturbation propagation over the declared horizon T?	The matrix exponential combines all couplings nonlinearly and may be dominated by rare spurious unstable modes; median and tail statistics should accompany the mean.
Directional E_Φ	What is the finite-time consequence of perturbing a specific physical state—for example lateral, yaw, roll, or wheel-speed perturbations?	Direction choice is part of the validation definition and must be fixed before interpreting the result.
I_J	Where does the reference Jacobian vary strongly, indicating rapid change of the local tangent map?	Useful for locating derivative complexity, but it is not itself a prediction error and does not guarantee finite-time relevance.
Symmetric growth	Where can a perturbation experience strong instantaneous longitudinal amplification or contraction?	An instantaneous directional quantity; it does not represent transverse reorientation or complete finite-time behavior.
g_F^C	Where does the vector field strongly reorient a perturbation transverse to its current direction?	Mechanism-specific and direction dependent; it is not a replacement for the full Jacobian or propagator.
Non-normality	Where can non-orthogonal local modes permit transient amplification beyond eigenvalue intuition?	High non-normality is only a susceptibility indicator. In the present post-hoc analysis the M1 propaga-

Table 15 Summary: which element of training each process changes, its global effect, and its main state-level effect. Evidence: C = confirmatory aggregate, E = post-hoc exploratory (ten seeds), D = descriptive (three seeds), — = not evaluated.

Training process	Data	Objective	Representation	Global effect	Main state-level effect	Evid.
Random / coverage	✓			coverage: E_Φ −20%, $E_{A,\mathrm{rel}}$ n.s.	better attitude/wheel propagation, no equation change	E
Physics selection (Exp. 2)	✓			$E_{A,\mathrm{rel}}$ −2.1%, E_Φ n.s.	v_y row; yaw/roll/pitch propagation	C/E
M1	✓			$E_{A,\mathrm{rel}}$ −16.5%, E_Φ +27.6%	v_x, pitch-rate, wheels better; v_y, roll rate worse	C/E
M2	✓			$E_{A,\mathrm{rel}}$ −2.6%, E_Φ +8.6%	roll/pitch influences, yaw/roll/wind propagation worse	C/E
DualWeight		✓		no matched gain	negligible	C/E
LRA		✓		E_Φ +8.5%	v_y, roll-rate rows and all directions worse	C/E
GradNorm		✓		neutral	negligible	C/E
SoftAdapt		✓		no matched gain	roll propagation worse (n.s.)	C/E
ReLoBRaLo		✓		no matched gain	body-mode frequency errors worse	C/E
Adam→L-BFGS-B		✓		$E_{A,\mathrm{rel}}$ +0.5%, E_Φ n.s.	spurious attitude/heave couplings; yaw/roll vs pitch/heave trade	C/E
Analytical prior $I1$			✓	small $E_{A,\mathrm{rel}}$, median E_Φ better at 10%	rotating-frame entries; benefit large only	D

Table A1 Modal and stability effect (relative change, %, and symbol against the matched comparator). $|\Delta\alpha|$: spectral-abscissa error (comparator value in parentheses, 1/s); stab.: stability-class preservation at $\tau_{\text{stab}} = 0.1\,\text{s}^{-1}$ (positive = better); $|\Delta\omega_n|$, $|\Delta\zeta|$: matched-mode natural-frequency and damping-ratio errors of local modes dominated by the stated group in the declared normalized basis (diagnostic labels, not invariant vehicle modes).

			$\lvert\Delta\omega_n\rvert$					$\lvert\Delta\zeta\rvert$				
Training process	$\lvert\Delta\alpha\rvert$	stab.	lat.	yaw	roll	pitch	wheels	lat.	yaw	roll	pitch	wheels
A: det. coverage vs random	-35 ++ (0.47)	+17 (++)	+8 (−)	-14 (++)	-24 ++	-14 (++)	-43 ++	+4 0	+1 0	-1 0	-5 (+)	-4 0
B: physics family vs agnostic	-16 ++ (0.49)	+15 ++	+7 (−)	+13 (−−)	-67 ++	-35 ++	-8 (+)	-8 (+)	+13 −−	-17 ++	-23 ++	-27 ++
C: physics+coverage vs random	-35 ++ (0.47)	+21 (++)	-3 0	-1 0	-18 (++)	-7 (+)	-5 0	-25 (++)	-11 (++)	-10 (++)	+1 0	-24 ++
C: physics+coverage vs det. cov.	+1 0 (0.29)	+4 0	-10 (+)	+22 (−−)	+8 (−)	+10 (−)	+63 −−	-24 (++)	-11 (++)	-9 (+)	+8 (−)	-19 ++
D: M1 Lagrangian	+64 −− (0.29)	-17 −−	+250 −−	+181 (−−)	+22 (−−)	+1 0	+103 (−−)	+126 −−	+146 −−	+20 −−	+12 (−−)	+16 (−−)
E: M2 primal–dual	+30 −− (0.29)	-11 (−−)	+19 (−−)	+8 (−)	+9 (−)	+7 (−)	+1 0	+21 (−−)	+11 (−−)	+20 −−	+11 (−−)	+29 −−
F: DualWeight	+5 (−) (0.29)	+2 0	+44 (−−)	+30 (−−)	+28 (−−)	+14 (−−)	+53 (−−)	+31 (−−)	-0 0	+9 (−)	+0 0	+10 (−−)
F: LRA	+17 (−−) (0.29)	-10 (−)	+19 (−−)	+4 0	+12 (−−)	+0 0	-30 ++	+57 (−−)	+35 −−	+22 −−	+28 −−	+37 −−
F: GradNorm	+2 0 (0.29)	+1 0	+38 (−−)	+27 (−−)	+8 (−)	+10 (−−)	+19 (−−)	+53 (−−)	+11 (−−)	+6 (−)	+5 0	+4 0
F: SoftAdapt	+9 (−) (0.29)	+3 0	+28 (−−)	+28 (−−)	+21 (−−)	+8 (−)	+15 (−−)	+41 (−−)	+14 (−−)	+14 −−	+7 (−)	+12 −−
F: ReLoBRaLo	+3 0 (0.29)	+4 0	+46 (−−)	+29 −−	+24 −−	+16 −−	+32 (−−)	+47 (−−)	+9 (−)	+17 −−	+9 (−)	+13 (−−)
G: Adam→L-BFGS-B (P+C set)	+33 (−−) (0.29)	-13 (−−)	+153 −−	+130 −−	+164 −−	+32 −−	+367 −−	+40 (−−)	+16 (−−)	-9 (+)	+21 (−−)	-26 ++
G: Adam→L-BFGS-B (O1 median)	+16 −− (0.32)	-6 (−)	+76 −−	+109 −−	+120 −−	+38 −−	+310 −−	+4 0	-1 0	-13 ++	+21 −−	-30 ++
H: *I*1 prior, A0, 10%	-52 (++) (0.09)	+7 (+)	-25 (++)	-16 (++)	+13 (−−)	-3 0	+61 (−−)	-18 (++)	+20 (−−)	-9 (+)	+1 0	-30 (++)
H: *I*1 prior, A0, 2%	-13 (++) (0.58)	+7 (+)	-14 (++)	+2 0	+36 (−−)	+8 (−)	-30 (++)	+31 (−−)	-6 (+)	+1 0	+1 0	-24 (++)
H: *I*1 prior, plain64, 10%	-19 (++) (0.08)	+2 0	-27 (++)	-42 (++)	-34 (++)	-19 (++)	-0 0	-32 (++)	-62 (++)	-31 (++)	-43 (++)	-23 (++)
I: no affine skip, 10%	-2 0 (0.09)	+1 0	+231 (−−)	+113 (−−)	+173 (−−)	+244 (−−)	+730 (−−)	+75 (−−)	+141 (−−)	+33 (−−)	+228 (−−)	+46 (−−)
I: medium256, 10%	+566 (−−) (0.09)	-10 (−)	+48 (−−)	-9 (+)	-18 (++)	-19 (++)	+188 (−−)	+56 (−−)	+38 (−−)	-32 (++)	+14 (−−)	-29 (++)
I: medium256, 2%	+20 (−−) (0.58)	+23 (++)	+8 (−)	-24 (++)	+60 (−−)	+15 (−−)	+218 (−−)	+3 0	+29 (−−)	+10 (−)	+75 (−−)	-37 (++)
I: width256, 10%	+32 (−−) (0.09)	-1 0	+165 (−−)	+11 (−−)	+88 (−−)	-10 (++)	+964 (−−)	+13 (−−)	+23 (−−)	-30 (++)	+2 0	-51 (++)
I: depth6, 10%	+42 (−−) (0.09)	-3 0	+128 (−−)	+26 (−−)	+112 (−−)	+48 (−−)	+1110 (−−)	+36 (−−)	+65 (−−)	+37 (−−)	+47 (−−)	-3 0
I: SiLU, 10%	+17 (−−) (0.09)	+3 0	+23 (−−)	-12 (++)	+12 (−−)	+8 (−)	-32 (++)	+28 (−−)	+4 0	+29 (−−)	+9 (−)	+48 (−−)
I: ReLU, 10%	+151 (−−) (0.09)	-7 (−)	+22 (−−)	+33 (−−)	+104 (−−)	+80 (−−)	+54 (−−)	+84 (−−)	+69 (−−)	+54 (−−)	+88 (−−)	+31 (−−)

Appendix A State-resolved supplementary tables

This appendix reports the complete state-resolved results summarized in Sec. 6.10. All entries are post-hoc: v5/v6 rows are exploratory (ten seeds), Experiment 3 rows descriptive (three seeds); symbols follow Table 10. Complete CSV files with comparator and intervention values, confidence intervals, raw and Holm-adjusted p-values for every intervention, state, group, block, entry, and mode are archived with the analysis code.

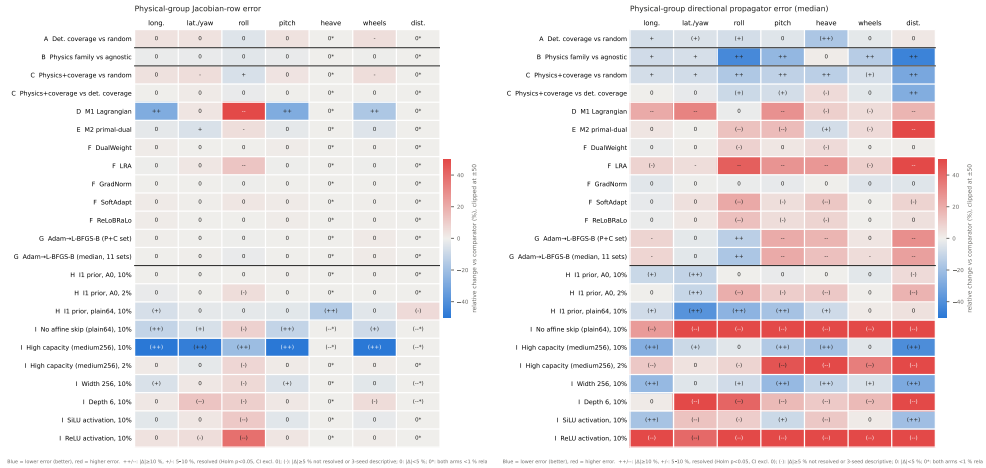


Fig. A1 Physical-group version of Fig. 1: group Jacobian-row error (left) and group directional propagator error, pointwise median (right). Groups: longitudinal $\{v_x\}$, lateral/yaw $\{v_y, r\}$, roll $\{\phi, \dot{\phi}\}$, pitch $\{\theta_p, \dot{\theta}_p\}$, heave $\{z, \dot{z}\}$, wheels, disturbance $\{F_{\text{wind}}\}$. Heave and disturbance rows are linear and exactly represented by every affine-skip learner (0*).

Table A2 Grouped Jacobian-block error (mean absolute block error in scaled coordinates): relative change (%) against the matched comparator. Block $g \leftarrow h$ is the sensitivity of the derivatives of group g to the states of group h. The first four columns are the large blocks that dominate the Frobenius $E_{A,\mathrm{rel}}$ numerator. With 64 blocks in the Holm family no block can be resolved with ten seeds, so all block entries are descriptive; the M1 values reproduce the frozen post-hoc block analysis exactly.

	pitch←long.	wheels←long.	pitch←wheels	wheels←wheels	roll←long.	lat.←long.	roll←wheels	lat.←wheels	yaw←lat.	lat.←yaw	roll←roll
B: physics family vs agnostic	-1	-1	-2	-1	+27	+2	+2	-1	-15	-8	-14
C: physics+coverage vs random	+2	+1	+2	+2	+9	+30	+3	+5	-10	+1	-24
D: M1 Lagrangian	-10	-6	-7	-2	+152	+100	+79	+76	+6	+31	+41
E: M2 primal–dual	-1	-1	-1	-2	-0	-17	+1	-4	+5	-7	+25
F: LRA	-0	-0	+0	-0	-3	-2	+3	-0	+21	+12	+20
F: SoftAdapt	+1	+0	+0	+0	+8	+5	+2	+2	-0	+4	+11
G: Adam→L-BFGS-B (P+C set)	-1	-1	-0	+0	-3	-6	+5	+5	+1	-8	+2
G: Adam→L-BFGS-B (O1 median)	-1	-1	-0	+0	-3	-5	+5	+5	+2	-7	-1

Table A3 State-to-state propagation $|\widehat{\Phi}_{ij} - \Phi_{ij}|$ (pointwise median, $T = 0.5\,\mathrm{s}$) for entries declared before computation; $j \to i$ denotes the response of state i to a perturbation of state j. Relative change (%) and symbol against the matched comparator.

	$v_y \to r$	$r \to v_y$	$r \to \phi$	$\phi \to v_y$	$v_y \to \phi$	$\omega_{FL} \to v_x$	$\omega_{RL} \to v_x$	$\omega_{FL} \to \phi$	$v_x \to \theta_p$	$v_x \to \omega_{FL}$	$F_\mathrm{w} \to v_y$
A: det. coverage vs random	-21 (++)	-14 (++)	-22 (++)	-4 0	-2 0	-25 (++)	-21 (++)	-26 (++)	+4 0	+8 (−)	+20 (−−)
B: physics family vs agnostic	-32 ++	-49 ++	-22 ++	-61 ++	-32 ++	-6 (+)	-3 0	-11 (++)	-17 ++	-17 ++	-57 ++
C: physics+coverage vs random	-31 ++	-25 (++)	-31 ++	-17 (++)	-28 ++	-6 (+)	-8 (+)	-24 (++)	+12 (−−)	-8 (+)	-30 ++
C: physics+coverage vs det. cov.	-9 (+)	-7 (+)	-5 (+)	-10 (+)	-23 (++)	+24 −−	+16 −−	+1 0	+8 (−)	-14 ++	-40 ++
D: M1 Lagrangian	+68 −−	-4 0	+23 −−	-52 ++	+35 (−−)	-19 ++	+0 0	+43 −−	-8 (+)	+22 −−	+1 0
E: M2 primal–dual	+9 (−)	+23 −−	+9 (−)	+26 (−−)	+14 (−−)	+5 (−)	+3 0	+21 (−−)	+2 0	+7 (−)	+73 −−
F: DualWeight	-1 0	+1 0	+16 (−−)	+7 (−)	+5 0	-2 0	-3 0	+11 (−−)	-5 (+)	-4 0	+11 (−−)
F: LRA	+38 −−	+69 −−	+73 −−	+50 −−	+43 −−	+9 (−)	+5 (−)	+86 −−	-4 0	+9 (−)	+132 −−
F: GradNorm	+1 0	+2 0	-3 0	-9 (+)	+6 (−)	-3 0	-3 0	+6 (−)	-2 0	-5 0	-4 0
F: SoftAdapt	+16 (−−)	+6 (−)	+32 (−−)	+9 (−)	+19 (−−)	+3 0	+3 0	+24 −−	-8 (+)	-8 (+)	+16 (−−)
F: ReLoBRaLo	-1 0	+8 (−)	+23 (−−)	+7 (−)	+1 0	+1 0	+3 0	+22 (−−)	-3 0	-5 0	+9 (−)
G: Adam→L-BFGS-B (P+C set)	-12 (++)	+9 (−)	-20 (++)	-6 (+)	-19 (++)	+4 0	+6 (−)	-4 0	+10 (−−)	+60 −−	+20 (−−)
G: Adam→L-BFGS-B (O1 median)	-15 ++	-8 (+)	-24 ++	-16 (++)	-18 ++	+7 (−)	+6 (−)	-6 (+)	+15 −−	+61 −−	+13 (−−)
H: *I*1 prior, A0, 10%	-28 (++)	-13 (++)	+13 (−−)	-12 (++)	+3 0	-18 (++)	-17 (++)	+8 (−)	-7 (+)	-9 (+)	+13 (−−)
H: *I*1 prior, A0, 2%	+25 (−−)	-10 (++)	+33 (−−)	+14 (−−)	+29 (−−)	-15 (++)	-32 (++)	+20 (−−)	-8 (+)	+9 (−)	+27 (−−)
H: *I*1 prior, plain64, 10%	-39 (++)	-57 (++)	-33 (++)	-34 (++)	-36 (++)	+3 0	-12 (++)	-38 (++)	+30 (−−)	+7 (−)	-31 (++)
I: no affine skip, 10%	+85 (−−)	+72 (−−)	+63 (−−)	+66 (−−)	+21 (−−)	-12 (++)	+23 (−−)	+68 (−−)	+158 (−−)	-51 (++)	+24 (−−)
I: medium256, 10%	-37 (++)	-36 (++)	-11 (++)	-27 (++)	-32 (++)	-4 0	+42 (−−)	-16 (++)	-23 (++)	-67 (++)	-42 (++)
I: medium256, 2%	+3 0	-29 (++)	-37 (++)	-9 (+)	+58 (−−)	-15 (++)	-17 (++)	-17 (++)	+5 0	+41 (−−)	+68 (−−)
I: width256, 10%	-45 (++)	-15 (++)	-10 (+)	-54 (++)	-13 (++)	-6 (+)	+26 (−−)	-18 (++)	-0 0	-45 (++)	-43 (++)
I: depth6, 10%	+34 (−−)	-3 0	+57 (−−)	+63 (−−)	+4 0	+41 (−−)	+49 (−−)	+15 (−−)	+11 (−−)	+1 0	+81 (−−)
I: SiLU, 10%	+59 (−−)	+97 (−−)	+60 (−−)	+178 (−−)	-0 0	+82 (−−)	+64 (−−)	+31 (−−)	-8 (+)	-19 (++)	+17 (−−)
I: ReLU, 10%	+69 (−−)	+256 (−−)	+225 (−−)	+295 (−−)	+26 (−−)	+35 (−−)	+38 (−−)	+90 (−−)	+22 (−−)	+8 (−)	+408 (−−)

Table A4 State-wise nominal vector-field error $E_{F,i} = \mathrm{mean}_q|\widehat{\dot{F}}_i - \bar{F}_i|$ (normalized coordinates): relative change (%) and symbol against the matched comparator.

Intervention	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_w
A: det. coverage vs random	-4 0	-12 (++)	-1 0	0*	-18 (++)	0*	-0 0	++*	++*	+17 (−−)	-2 0	-13 (++)	-8 (+)	0*
B: physics family vs agnostic	-21 ++	+5 (−)	-15 ++	0*	-1 0	0*	-25 ++	0*	0*	-23 ++	-10 (+)	-5 (+)	-35 ++	0*
C: physics+coverage vs random	+5 0	+5 0	+33 (−−)	0*	-5 (+)	0*	+11 (−−)	++*	++*	+4 0	-4 0	+26 (−−)	-24 (++)	0*
C: physics+coverage vs det. cov.	+13 (−−)	+20 (−−)	+33 (−−)	0*	+18 (−−)	0*	+15 (−−)	0*	0*	-7 (+)	+4 0	+44 −−	-18 (++)	0*
D: M1 Lagrangian	-8 (+)	+72 (−−)	-0 0	0*	+116 −−	0*	-15 ++	−−*	0*	-5 0	+13 (−−)	-13 ++	+43 −−	0*
E: M2 primal–dual	+5 0	+26 (−−)	-21 (++)	0*	+53 (−−)	0*	-5 (+)	0*	0*	-12 (++)	-15 (++)	-17 (++)	+50 −−	0*
F: DualWeight	+13 (−−)	-2 0	+0 0	0*	-2 0	0*	+10 (−−)	−−*	0*	+11 (−−)	+16 (−−)	+6 (−)	-7 (+)	0*
F: LRA	-6 (+)	-16 (++)	+1 0	0*	-11 (++)	0*	-6 (+)	−−*	0*	-1 0	+3 0	-6 (+)	+13 (−−)	0*
F: GradNorm	+5 0	-2 0	+2 0	0*	+6 (−)	0*	+3 0	0*	−−*	+3 0	+10 (−−)	+1 0	+0 0	0*
F: SoftAdapt	+9 (−)	-10 (+)	-6 (+)	0*	-3 0	0*	+8 (−)	0*	0*	+7 (−)	+5 (−)	+8 (−)	-6 (+)	0*
F: ReLoBRaLo	+13 (−−)	-14 (++)	+1 0	0*	-10 (+)	0*	+11 (−−)	0*	0*	+12 (−−)	+18 (−−)	+7 (−)	-3 0	0*
G: Adam→L-BFGS-B (P+C set)	+22 −−	-5 0	-0 0	0*	+1 0	0*	+16 −−	0*	0*	+18 (−−)	+43 −−	+27 (−−)	+106 −−	0*
G: Adam→L-BFGS-B (O1 median)	+19 −−	-4 0	+9 (−)	0*	+1 0	0*	+15 −−	0*	0*	+14 (−−)	+38 −−	+26 −−	+91 −−	0*
H: *I*1 prior, A0, 10%	-1 0	-30 (++)	+1 0	0*	+1 0	0*	+3 0	0*	0*	-0 0	+4 0	-2 0	+1 0	0*
H: *I*1 prior, A0, 2%	-1 0	-13 (++)	+8 (−)	0*	+6 (−)	0*	+7 (−)	0*	0*	+3 0	+4 0	+7 (−)	+3 0	0*
H: *I*1 prior, plain64, 10%	-8 (+)	-37 (++)	-5 (+)	(++*)	+2 0	(++*)	-2 0	(++*)	+4 0	-1 0	-2 0	-2 0	-3 0	+7 (−)
I: no affine skip, 10%	-17 (++)	+3 0	-14 (++)	(−−*)	+9 (−)	(−−*)	-17 (++)	(−−*)	(−−*)	-26 (++)	-16 (++)	-21 (++)	-3 0	(−−*)
I: medium256, 10%	-68 (++)	-35 (++)	-57 (++)	0*	-37 (++)	0*	-70 (++)	0*	0*	-73 (++)	-64 (++)	-77 (++)	-61 (++)	0*
I: medium256, 2%	+1 0	+1 0	+7 (−)	(−−*)	+11 (−−)	0*	+4 0	0*	(−−*)	+2 0	+11 (−−)	+10 (−)	+18 (−−)	0*
I: width256, 10%	-7 (+)	-14 (++)	-5 (+)	0*	-3 0	0*	-9 (+)	0*	0*	-24 (++)	+1 0	-17 (++)	+4 0	0*
I: depth6, 10%	-3 0	-2 0	+17 (−−)	(−−*)	+11 (−−)	0*	-6 (+)	0*	0*	-4 0	+2 0	+2 0	+12 (−−)	0*
I: SiLU, 10%	-9 (+)	-21 (++)	-2 0	0*	-2 0	0*	-7 (+)	0*	0*	-11 (++)	-3 0	-5 0	+2 0	0*
I: ReLU, 10%	+2 0	+44 (−−)	+21 (−−)	0*	+75 (−−)	0*	+4 0	0*	0*	-1 0	+3 0	-0 0	+4 0	0*

Table A5 State-wise Jacobian-row error $E_{A,\mathrm{row}}(i)$ (accuracy of the local differential equation of state i): relative change (%) and symbol.

Intervention	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
A: det. coverage vs random	+40	-20	+50	0*	-40	0*	+40	0*	0*	+20	+6 −	+40	+10 −	0*
B: physics family vs agnostic	-20	-10 ++	-10	0*	-40	0*	-20	0*	0*	+20	-40	+40	-30	0*
C: physics+coverage vs random	+40	-40	+8 −	0*	-6 +	0*	+40	0*	0*	+50	+6 −	+6 −	+8 −	0*
C: physics+coverage vs det. cov.	+00	-20	+30	0*	-20	0*	-00	0*	0*	+30	+00	+20	-20	0*
D: M1 Lagrangian	-28 ++	+44 −−	+30	0*	+64 −−	0*	-28 ++	0*	0*	-19 ++	-23 ++	-30	-21 ++	0*
E: M2 primal–dual	-30	+20	-8 +	0*	+5 (−)	0*	-30	0*	0*	-9 +	-40	-40	-40	0*
F: DualWeight	+10	-10	-00	0*	-00	0*	+10	0*	0*	+10	+00	+10	+00	0*
F: LRA	-10	+14 −−	+20	0*	+12 −−	0*	-10	0*	0*	-10	-00	-10	-00	0*
F: GradNorm	+00	-00	-00	0*	-10	0*	+00	0*	0*	+00	+00	+00	+00	0*
F: SoftAdapt	+20	+20	+00	0*	+20	0*	+10	0*	0*	+10	+00	+10	+10	0*
F: ReLoBRaLo	+10	-00	+00	0*	-00	0*	+10	0*	0*	+10	+00	+10	+00	0*
G: Adam→L-BFGS-B (P+C set)	+00	-10	+00	0*	-00	0*	+00	0*	0*	+20	-10	+20	-10	0*
G: Adam→L-BFGS-B (O1 median)	-00	-20	+00	0*	-20	0*	-00	0*	0*	+10	-10	+20	-10	0*
H: *I*1 prior, A0, 10%	+00	-7 (+)	-00	0*	-00	0*	+10	0*	0*	+30	-00	+20	+00	0*
H: *I*1 prior, A0, 2%	-20	-20	-00	0*	+5 (−)	0*	+00	0*	0*	+00	+00	+00	-00	0*
H: *I*1 prior, plain64, 10%	-6 (+)	-15 (++)	-20	(++*)	-20	(++*)	-30	(++*)	+17 (−−)	-20	-7 (+)	-30	-6 (+)	+6 (−)
I: no affine skip, 10%	-11 (++)	+9 (−)	-7 (+)	(−−*)	+9 (−)	(−−*)	-11 (++)	(−−*)	(−−*)	-20 (++)	-10 (++)	-11 (++)	-40	(−−*)
I: medium256, 10%	-57 (++)	-14 (++)	-47 (++)	0*	-21 (++)	(−−*)	-58 (++)	0*	(−−*)	-67 (++)	-60 (++)	-70 (++)	-58 (++)	(−−*)
I: medium256, 2%	+20	+7 (−)	-30	0*	+7 (−)	0*	+10	0*	0*	+10	+20	+10	+00	0*
I: width256, 10%	-6 (+)	+40	+20	0*	+6 (−)	0*	-5 (+)	0*	0*	-16 (++)	-30	-9 (+)	-20	(−−*)
I: depth6, 10%	+20	+16 (−−)	+13 (−−)	0*	+9 (−)	0*	+20	0*	0*	-7 (+)	+7 (−)	+10	+7 (−)	(−−*)
I: SiLU, 10%	-30	+40	+20	0*	+11 (−−)	0*	-40	0*	0*	-12 (++)	-10	-5 (+)	+30	0*
I: ReLU, 10%	+40	+42 (−−)	+50	0*	+38 (−−)	0*	+30	0*	0*	-00	+30	-30	+50	0*

Table A6 State-wise Jacobian-column error $E_{A,\mathrm{col}}(j)$ (accuracy of the influence of state j on the rest of the vehicle): relative change (%) and symbol.

Intervention	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
A: det. coverage vs random	+40	+30	-16 ++	-16 ++	-26 ++	-22 ++	-14 (++)	-12 (++)	-20	+00	+7 −	+30	+9 −	-8 (+)
B: physics family vs agnostic	-10	+29 −−	+6 −	-8 +	-13 ++	-20 ++	-30 ++	-26 ++	-28 ++	+30	-6 +	+10 −−	-7 +	-42 ++
C: physics+coverage vs random	+30	+13 (−−)	-12 ++	-18 ++	-29 ++	-23 ++	-24 ++	-19 ++	-15 ++	+50	+8 −	+20	+7 −	-33 ++
C: physics+coverage vs det. cov.	-00	+12 (−−)	+40	-20	-30	+10	-12 (++)	-7 (+)	-13 ++	+40	+10	-10	-20	-27 ++
D: M1 Lagrangian	-28 ++	+78 −−	+10 −	+17 −−	+40 −−	+22 (−−)	+18 (−−)	+10	+36 −−	-12 ++	-26 ++	+10	-22 ++	-40
E: M2 primal–dual	-20	-18 ++	+9 −	+14 −−	+32 −−	+58 −−	+25 (−−)	+30	+15 −−	-10 +	-40	-00	-50	+60 −−
F: DualWeight	+10	+30	-00	+10	+20	-20	+7 (−)	+50	-10	+00	-00	+00	+00	-5 (+)
F: LRA	-10	-5 (+)	+7 (−)	+9 −	+22 −−	+58 −−	+36 −−	+16 −−	+29 −−	+00	-00	+00	+00	+76 −−
F: GradNorm	+00	-30	-20	-40	-00	+10	+30	+00	-10	+10	-00	+00	+00	-00
F: SoftAdapt	+10	+10	+00	+50	+13 (−−)	+10	+13 (−−)	+12 (−−)	+20	+10	-00	+00	-00	-7 (+)
F: ReLoBRaLo	+10	-40	+10	+10	+8 (−)	-10	+9 (−)	+8 (−)	+20	+00	-00	+00	-00	-10 (+)
G: Adam→L-BFGS-B (P+C set)	-10	+39 −−	+61 −−	+72 −−	+107 −−	+170 −−	+171 −−	+157 −−	+125 −−	-10	-00	+10	+00	+160 −−
G: Adam→L-BFGS-B (O1 median)	-10	+37 −−	+56 −−	+66 −−	+93 −−	+156 −−	+151 −−	+132 −−	+107 −−	-10	-10	+00	-10	+162 −−
H: *I*1 prior, A0, 10%	+10	-40	-21 (++)	+13 (−−)	+18 (−−)	+12 (−−)	+24 (−−)	+19 (−−)	+39 (−−)	+10	-10	+30	-00	+10 (−−)
H: *I*1 prior, A0, 2%	+00	+20	-7 (+)	-10	+18 (−−)	-10 (+)	+8 (−)	-20	-7 (+)	+00	+00	+00	+00	+17 (−−)
H: *I*1 prior, plain64, 10%	-40	+30	-28 (++)	-30	-17 (++)	+16 (−−)	-35 (++)	+15 (−−)	-17 (++)	-20	-6 (+)	-40	-5 (+)	+18 (−−)
I: no affine skip, 10%	-12 (++)	-16 (++)	-30	-30	+18 (−−)	+6 (−)	+21 (−−)	+20	+63 (−−)	-20 (++)	-14 (++)	-32 (++)	-8 (+)	-30 (++)
I: medium256, 10%	-60 (++)	-32 (++)	-23 (++)	-18 (++)	-19 (++)	-27 (++)	-38 (++)	-43 (++)	-47 (++)	-55 (++)	-67 (++)	-55 (++)	-57 (++)	-73 (++)
I: medium256, 2%	+10	+23 (−−)	+20 (−−)	+7 (−)	+40 (−−)	+15 (−−)	+133 (−−)	+59 (−−)	+20 (−−)	-00	+00	+10	+10	+64 (−−)
I: width256, 10%	-7 (+)	-10	+5 (−)	+9 (−)	+8 (−)	+11 (−−)	-20	+6 (−)	+7 (−)	-21 (++)	-6 (+)	-23 (++)	-30	-32 (++)
I: depth6, 10%	-50	+102 (−−)	+115 (−−)	+85 (−−)	+93 (−−)	+69 (−−)	+46 (−−)	+48 (−−)	+33 (−−)	-8 (+)	-20	-15 (++)	-40	+75 (−−)
I: SiLU, 10%	-40	-12 (++)	-10	+12 (−−)	+5 (−)	-15 (++)	-32 (++)	-12 (++)	+10 (−−)	-11 (++)	-00	-11 (++)	+30	-50 (++)
I: ReLU, 10%	+20	+36 (−−)	+40 (−−)	+29 (−−)	+64 (−−)	+72 (−−)	+96 (−−)	+88 (−−)	+78 (−−)	+00	+20	-8 (+)	+50	+66 (−−)

Table A7 Directional propagator error $E_{\Phi,\mathrm{dir}}(j)$ for the canonical perturbation $\delta x_0 = e_j$ (pointwise median over operating points): relative change (%) and symbol.

Intervention	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
A: det. coverage vs random	-7 (+)	-6 (+)	-21 (++)	-17 (++)	-13 (++)	-6 (+)	-40	-28 ++	-40	-22 ++	-14 (++)	-24 ++	-13 (++)	+20
B: physics family vs agnostic	-8 +	-12 ++	-35 ++	-40 ++	-36 ++	-27 ++	-22 ++	-30	-20	-17 ++	-20 ++	-30	-12 ++	-45 ++
C: physics+coverage vs random	-7 +	-10 ++	-29 ++	-22 ++	-20 ++	-14 ++	-19 ++	-19 ++	-9 +	-16 ++	-19 ++	-17 ++	-19 ++	-30 ++
C: physics+coverage vs det. cov.	+00	-40	-10 (+)	-6 (+)	-8 (+)	-7 (+)	-15 ++	+14 (−−)	-40	+7 (−)	-6 (+)	+8 −	-7 (+)	-28 ++
D: M1 Lagrangian	+21 −−	+24 −−	+7 (−)	+20	+12 (−−)	+24 −−	+37 −−	+11 (−−)	+9 (−)	+10 (−−)	+30	+6 (−)	+20 (−−)	+18 (−−)
E: M2 primal–dual	+30	+40	+15 −−	+22 −−	+24 −−	+17 (−−)	+20 −−	-10 (++)	-30	+7 (−)	+20 −−	+6 (−)	+29 −−	+59 −−
F: DualWeight	-10	+20	+00	+10 (−−)	+12 (−−)	+10	+8 (−)	+8 (−)	+30	-10	+00	-10	+30	+10
F: LRA	+8 −	+12 −−	+41 −−	+40	+37 −−	+34 −−	+26 −−	+48 −−	+13 −−	+21 −−	+18 −−	+14 −−	+13 −−	+98 −−
F: GradNorm	-40	-00	-10	+10	+10	+00	+10	-20	-10	-30	-10	-40	-00	-40
F: SoftAdapt	-20	+8 (−)	+9 (−)	+23 (−−)	+23 (−−)	+6 (−)	+7 (−)	+17 (−−)	+6 (−)	+6 (−)	+5 (−)	+30	+40	+8 (−)
F: ReLoBRaLo	-10	+30	+7 (−)	+14 (−−)	+14 (−−)	+5 (−)	+8 (−)	+11 (−−)	+30	-00	+30	-10	+20	+40
G: Adam→L-BFGS-B (P+C set)	+8 (−)	-40	-9 +	-14 ++	-10 (+)	+23 −−	+21 −−	+24 −−	+11 (−−)	+50	+40	+50	+7 (−)	+29 −−
G: Adam→L-BFGS-B (O1 median)	+8 −	-6 +	-15 ++	-21 ++	-12 ++	+21 −−	+16 −−	+23 −−	+9 −	+40	-10	+6 −	-10	+24 −−
H: *I*1 prior, A0, 10%	-9 (+)	-10 (+)	+00	-10	-20	+20	-6 (+)	-50	-30	-8 (+)	-7 (+)	+10	-40	+6 (−)
H: *I*1 prior, A0, 2%	+00	-00	+7 (−)	+24 (−−)	+5 (−)	+17 (−−)	+10 (−−)	+20	+16 (−−)	+00	+10	-30	-10	+18 (−−)
H: *I*1 prior, plain64, 10%	-8 (+)	-29 (++)	-31 (++)	-26 (++)	-23 (++)	-26 (++)	-15 (++)	-15 (++)	-10 (++)	-12 (++)	-12 (++)	-8 (+)	-20	+40
I: no affine skip, 10%	+25 (−−)	+39 (−−)	+52 (−−)	+49 (−−)	+67 (−−)	+109 (−−)	+81 (−−)	+77 (−−)	+62 (−−)	+61 (−−)	+54 (−−)	+57 (−−)	+43 (−−)	+81 (−−)
I: medium256, 10%	-27 (++)	-19 (++)	-17 (++)	-8 (+)	-50	-12 (++)	-15 (++)	-24 (++)	-19 (++)	-9 (+)	-8 (+)	+7 (−)	-14 (++)	-41 (++)
I: medium256, 2%	+12 (−−)	-11 (++)	-33 (++)	-6 (+)	+42 (−−)	+60 (−−)	+47 (−−)	+70 (−−)	+69 (−−)	-15 (++)	+13 (−−)	-12 (++)	+10 (−)	+52 (−−)
I: width256, 10%	-22 (++)	-13 (++)	-8 (+)	-16 ++	-11 (+)	+23 −−	+21 −−	+24 −−	-7 (+)	-18 (++)	-30	-5 (+)	-6 (+)	-29 (++)
I: depth6, 10%	+10	+36 (−−)	+26 (−−)	+43 (−−)	+48 (−−)	+18 (−−)	+13 (−−)	+10 (−)	+7 (−)	+10 (−−)	+21 (−−)	+17 (−−)	+22 (−−)	+96 (−−)
I: SiLU, 10%	-18 (++)	+8 (−)	+32 (−−)	+8 (−)	+12 (−−)	-12 (++)	-50	-20	+8 (−)	+44 (−−)	+10 (−)	+41 (−−)	+8 (−)	-20 (++)
I: ReLU, 10%	+58 (−−)	+33 (−−)	+72 (−−)	+52 (−−)	+51 (−−)	+84 (−−)	+70 (−−)	+74 (−−)	+48 (−−)	+65 (−−)	+89 (−−)	+70 (−−)	+70 (−−)	+185 (−−)

Table A8 As Table A7 but with the pointwise mean; Experiment 3 means are dominated by rare exponential blow-ups and should be read with the median table.

Intervention	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
A: det. coverage vs random	-9 +	-27 ++	-37 ++	-27 ++	-30 ++	-6 (+)	-8 (+)	-36 ++	-11 (++)	-20 ++	-23 ++	-19 ++	-17 ++	-13 (++)
B: physics family vs agnostic	-6 (+)	-10	-22 ++	-22 ++	-22 ++	-28 ++	-25 ++	+30	+00	-14 ++	-18 ++	-9 +	-14 ++	-34 ++
C: physics+coverage vs random	-8 +	-26 ++	-35 ++	-21 ++	-30 ++	-16 ++	-18 ++	-24 ++	-14 ++	-17 ++	-24 ++	-16 ++	-19 ++	-35 ++
C: physics+coverage vs det. cov.	+00	+20	+40	+7 (−)	+10	-9 (+)	-11 ++	+22 (−−)	-30	+40	-10	+40	-30	-21 (++)
D: M1 Lagrangian	+21 −−	+52 −−	+6 (−)	+18 (−−)	+47 −−	+46 −−	+36 (−−)	+41 (−−)	+29 (−−)	+8 (−)	+7 (−)	+20	+12 (−−)	+39 −−
E: M2 primal–dual	+30	+10 −	+24 −−	+21 −−	+33 −−	+27 (−−)	+13 (−−)	-10 (++)	-10	+6 (−)	+16 −−	+40	+15 −−	+52 −−
F: DualWeight	-00	+6 (−)	-5 (+)	+20	+9 (−)	+40	+7 (−)	+14 (−−)	+50	+00	+30	-00	+30	+20
F: LRA	+50	+12 −−	+26 −−	+23 −−	+30 −−	+34 −−	+19 −−	+38 −−	+15 −−	+12 −−	+12 −−	+9 −	+9 −	+93 −−
F: GradNorm	-20	+20	-8 (+)	-6 (+)	+20	+00	-40	-10	+00	-20	-20	-20	-10	-40
F: SoftAdapt	-00	+16 (−−)	+30	+15 (−−)	+21 −−	+40	+8 (−)	+18 (−−)	+6 (−)	+40	+7 (−)	+20	+50	+6 (−)
F: ReLoBRaLo	-10	+8 (−)	-10	+8 (−)	+15 (−−)	+5 (−)	+7 (−)	+17 (−−)	+40	+10	+50	+10	+30	+20
G: Adam→L-BFGS-B (P+C set)	+9 −	+30	-20	-50	-00	+25 −−	+17 −−	+32 (−−)	+17 (−−)	+50	+5 (−)	+40	+40	+28 (−−)
G: Adam→L-BFGS-B (O1 median)	+8 −	-10	-7 +	-12 ++	-50	+25 −−	+17 −−	+27 −−	+14 −−	+40	-00	+40	-10	+28 −−
H: *I*1 prior, A0, 10%	-6 (+)	-12 (++)	-6 (+)	-5 (+)	-5 (+)	+30	-00	-30	+10	-30	-40	+00	-10	+12 (−−)
H: *I*1 prior, A0, 2%	+7 (−)	+12 (−−)	+15 (−−)	+24 (−−)	+37 (−−)	+8 (−)	+8 (−)	+30 (−−)	+16 (−−)	+16 (−−)	+12 (−−)	+13 (−−)	+9 (−)	+20 (−−)
H: *I*1 prior, plain64, 10%	+10	-31 (++)	-35 (++)	-38 (++)	-18 (++)	-26 (++)	-40	-16 (++)	-14 (++)	-7 (+)	-10 (+)	-7 (+)	-20	+15 (−−)
I: no affine skip, 10%	+16 (−−)	+38 (−−)	+51 (−−)	+68 (−−)	+64 (−−)	+132 (−−)	+89 (−−)	+86 (−−)	+78 (−−)	+32 (−−)	+37 (−−)	+21 (−−)	+21 (−−)	+91 (−−)
I: medium256, 10%	+6520 (−−)	+2317 (−−)	+1585 (−−)	+2741 (−−)	+1068 (−−)	+2606 (−−)	+870 (−−)	+1987 (−−)	+2103 (−−)	+33663 (−−)	+753 (−−)	+8275 (−−)	+2680 (−−)	+1689 (−−)
I: medium256, 2%	+25 (−−)	-5 (+)	-24 (++)	+11 (−−)	+22 (−−)	+42 (−−)	+79 (−−)	+52 (−−)	+69 (−−)	+8 (−)	-10	+9 (−)	-10	+58 (−−)
I: width256, 10%	-13 (++)	-17 (++)	-7 (+)	-24 (++)	-24 (++)	-15 (++)	-12 (++)	-10 (+)	-20	-6 (+)	-40	-20	-30	-26 (++)
I: depth6, 10%	+40	+39 (−−)	+25 (−−)	+66 (−−)	+51 (−−)	+23 (−−)	+17 (−−)	+12 (−−)	+12 (−−)	+7 (−)	+14 (−−)	+40	+12 (−−)	+100 (−−)
I: SiLU, 10%	-9 (+)	+30	+44 (−−)	+28 (−−)	+11 (−−)	-9 (+)	-10	+00	+7 (−)	+16 (−−)	+6 (−)	+11 (−−)	+5 (−)	-40
I: ReLU, 10%	+45 (−−)	+39 (−−)	+112 (−−)	+75 (−−)	+59 (−−)	+107 (−−)	+78 (−−)	+107 (−−)	+70 (−−)	+38 (−−)	+70 (−−)	+27 (−−)	+48 (−−)	+234 (−−)

Table A9 Jacobian-row error by state for the key interventions: comparator value, intervention value, absolute and relative difference.

	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
B: physics family vs agnostic														
comparator	3.28	0.713	1.66	8.35e-05	0.742	0.000133	2.72	8.41e-05	0.000149	9.4	8.56	8.55	9.08	0.000325
intervention	3.23	0.64	1.65	4.47e-05	0.714	0.000102	2.67	8.71e-05	8.93e-05	9.56	8.24	8.93	8.81	0.000427
abs. diff.	-0.05	-0.073	-0.01	-3.9e-05	-0.027	-3.1e-05	-0.05	+3e-06	-5.9e-05	+0.15	-0.32	+0.38	-0.27	+0.0001
rel. diff.	-2% 0	-10% ++	-1% 0	0*	-4% 0	0*	-2% 0	0*	0*	+2% 0	-4% 0	+4% 0	-3% 0	0*
D: M1 Lagrangian														
comparator	2.88	0.624	1.8	0.000142	0.648	0.00026	2.47	0.000184	0.000152	8.77	7.37	7.64	9.14	0.000463
intervention	2.07	0.9	1.85	0.000242	1.06	0.000191	1.78	0.00019	0.00015	7.09	5.67	7.39	7.27	0.000847
abs. diff.	-0.81	+0.28	+0.049	+0.0001	+0.41	-6.9e-05	-0.69	+5.7e-06	-2.1e-06	-1.7	-1.7	-0.25	-1.9	+0.00038
rel. diff.	-28% ++	+44% —	+3% 0	0*	+64% —	0*	-28% ++	0*	0*	-19% ++	-23% ++	-3% 0	-21% ++	0*
E: M2 primal–dual														
comparator	2.88	0.624	1.8	0.000142	0.648	0.00026	2.47	0.000184	0.000152	8.77	7.37	7.64	9.14	0.000463
intervention	2.8	0.633	1.65	4.51e-05	0.68	9.94e-05	2.41	9.47e-05	0.000299	8.02	7.08	7.33	8.78	0.00123
abs. diff.	-0.078	+0.0093	-0.14	-9.7e-05	+0.032	-0.00016	-0.064	-8.9e-05	+0.00015	-0.75	-0.29	-0.31	-0.36	+0.00076
rel. diff.	-3% 0	+2% 0	-8% +	0*	+5% (−)	0*	-3% 0	0*	0*	-9% +	-4% 0	-4% 0	-4% 0	0*
F: LRA														
comparator	2.88	0.624	1.8	0.000142	0.648	0.00026	2.47	0.000184	0.000152	8.77	7.37	7.64	9.14	0.000463
intervention	2.86	0.709	1.82	0.000182	0.725	7.33e-05	2.45	0.000177	0.000145	8.67	7.36	7.58	9.14	0.000197
abs. diff.	-0.016	+0.085	+0.027	+4e-05	+0.078	-0.00019	-0.016	-7.2e-06	-6.1e-06	-0.1	-0.012	-0.062	-0.0049	-0.00027
rel. diff.	-1% 0	+14% —	+2% 0	0*	+12% —	0*	-1% 0	0*	0*	-1% 0	-0% 0	-1% 0	-0% 0	0*
G: Adam→L-BFGS-B (O1 median)														
comparator	2.88	0.635	1.79	9.42e-05	0.668	0.000112	2.47	7.04e-05	9.29e-05	8.7	7.36	7.63	9.14	0.000211
intervention	2.88	0.624	1.8	0.000111	0.657	8.7e-05	2.47	6.38e-05	8.01e-05	8.77	7.3	7.79	9.06	0.00018
abs. diff.	-0.0047	-0.011	+0.0044	+1.7e-05	-0.011	-2.5e-05	-0.0071	-6.5e-06	-1.3e-05	+0.074	-0.052	+0.16	-0.082	-3.1e-05
rel. diff.	-0% 0	-2% 0	+0% 0	0*	-2% 0	0*	-0% 0	0*	0*	+1% 0	-1% 0	+2% 0	-1% 0	0*
H: $l1$ prior, A0, 10%														
comparator	1.01	0.299	0.834	0.000128	0.367	0.000628	0.949	5.82e-05	0.00023	4.85	2.06	5.15	2.9	0.00243
intervention	1.01	0.278	0.833	1.31e-05	0.366	6.47e-05	0.955	0.000212	0.000787	4.97	2.06	5.26	2.9	0.0015
abs. diff.	+0.0027	-0.02	-0.00059	-0.00011	-0.00078	-0.00056	+0.0059	+0.00015	+0.00056	+0.12	-0.0025	+0.1	+0.0013	-0.00093
rel. diff.	+0% 0	-7% (+)	-0% 0	0*	-0% 0	0*	+1% 0	0*	0*	+3% 0	-0% 0	+2% 0	+0% 0	0*
I: medium256, 10%														
comparator	1.01	0.299	0.834	0.000128	0.367	0.000628	0.949	5.82e-05	0.00023	4.85	2.06	5.15	2.9	0.00243
intervention	0.429	0.257	0.441	0.00488	0.29	0.00481	0.395	0.00395	0.0103	1.59	0.813	1.56	1.23	0.0125
abs. diff.	-0.58	-0.042	-0.39	+0.0048	-0.077	+0.0042	-0.55	+0.0039	+0.01	-3.3	-1.2	-3.6	-1.7	+0.01
rel. diff.	-57% (++)	-14% (++)	-47% (++)	0*	-21% (++)	(−−*)	-58% (++)	0*	(−−*)	-67% (++)	-60% (++)	-70% (++)	-58% (++)	(−−*)

Table A10 Directional propagator error (median) by perturbed state for the key interventions: comparator value, intervention value, absolute and relative difference.

	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
B: physics family vs agnostic														
comparator	0.525	1.21	1.69	3.51	3.56	0.651	0.368	0.84	0.934	0.718	0.744	0.823	0.81	0.582
intervention	0.482	1.07	1.09	2.1	2.26	0.473	0.286	0.811	0.915	0.592	0.59	0.796	0.709	0.318
abs. diff.	-0.043	-0.14	-0.6	-1.4	-1.3	-0.18	-0.082	-0.029	-0.018	-0.13	-0.15	-0.027	-0.1	-0.26
rel. diff.	-8% +	-12% ++	-35% ++	-40% ++	-36% ++	-27% ++	-22% ++	-3% 0	-2% 0	-17% ++	-20% ++	-3% 0	-12% ++	-45% ++
D: M1 Lagrangian														
comparator	0.452	1.08	1.12	2.27	2.13	0.424	0.238	0.622	0.814	0.519	0.541	0.629	0.608	0.281
intervention	0.543	1.34	1.2	2.26	2.35	0.523	0.325	0.685	0.881	0.571	0.556	0.663	0.724	0.331
abs. diff.	+0.091	+0.25	+0.076	-0.012	+0.22	+0.099	+0.087	+0.064	+0.067	+0.052	+0.014	+0.034	+0.12	+0.05
rel. diff.	+21% −−	+24% —	+7% (−)	+2% 0	+12% (−−)	+24% —	+37% —	+11% (—)	+9% (−)	+10% (−−)	+3% 0	+6% (−)	+20% (—)	+18% (—)
E: M2 primal–dual														
comparator	0.452	1.08	1.12	2.27	2.13	0.424	0.238	0.622	0.814	0.519	0.541	0.629	0.608	0.281
intervention	0.463	1.13	1.29	2.73	2.61	0.493	0.286	0.552	0.786	0.557	0.642	0.666	0.774	0.444
abs. diff.	+0.011	+0.043	+0.17	+0.46	+0.49	+0.069	+0.048	-0.069	-0.029	+0.037	+0.1	+0.037	+0.17	+0.16
rel. diff.	+3% 0	+4% 0	+15% —	+22% —	+24% —	+17% (−−)	+20% —	-10% (++)	-3% 0	+7% (−)	+20% —	+6% (−)	+29% —	+59% —
F: LRA														
comparator	0.452	1.08	1.12	2.27	2.13	0.424	0.238	0.622	0.814	0.519	0.541	0.629	0.608	0.281
intervention	0.487	1.22	1.58	3.14	2.89	0.562	0.301	0.906	0.916	0.626	0.636	0.713	0.684	0.549
abs. diff.	+0.035	+0.13	+0.45	+0.87	+0.76	+0.14	+0.063	+0.28	+0.1	+0.11	+0.094	+0.084	+0.077	+0.27
rel. diff.	+8% −	+12% —	+41% −−	+41% —	+37% —	+34% —	+26% —	+48% —	+13% —	+21% —	+18% —	+14% —	+13% —	+98% —
G: Adam→L-BFGS-B (O1 median)														
comparator	0.463	1.12	1.23	2.56	2.39	0.455	0.27	0.652	0.843	0.541	0.576	0.646	0.656	0.321
intervention	0.498	1.06	1.04	2.01	2.08	0.551	0.312	0.798	0.92	0.562	0.565	0.686	0.646	0.392
abs. diff.	+0.035	-0.065	-0.18	-0.55	-0.3	+0.097	+0.042	+0.15	+0.077	+0.021	-0.011	+0.04	-0.01	+0.071
rel. diff.	+8% −	-6% +	-15% ++	-21% ++	-12% ++	+21% —	+16% —	+23% —	+9% −	+4% 0	-1% 0	+6% −	-1% 0	+24% —
H: $l1$ prior, A0, 10%														
comparator	0.196	0.63	0.594	0.937	1.1	0.273	0.165	0.365	0.511	0.242	0.25	0.29	0.285	0.163
intervention	0.178	0.575	0.6	0.928	1.08	0.271	0.155	0.346	0.493	0.224	0.231	0.293	0.275	0.176
abs. diff.	-0.019	-0.055	+0.0058	-0.0092	-0.019	-0.0015	-0.01	-0.02	-0.018	-0.019	-0.018	+0.0029	-0.01	+0.013
rel. diff.	-9% (+)	-10% (+)	+0% 0	-1% 0	-2% 0	+2% 0	-6% (+)	-5% 0	-3% 0	-8% (+)	-7% (+)	+1% 0	-4% 0	+6% (−)
I: medium256, 10%														
comparator	0.196	0.63	0.594	0.937	1.1	0.273	0.165	0.365	0.511	0.242	0.25	0.29	0.285	0.163
intervention	0.143	0.506	0.479	0.865	1.04	0.234	0.138	0.274	0.412	0.218	0.229	0.311	0.245	0.0955
abs. diff.	-0.053	-0.12	-0.11	-0.072	-0.052	-0.039	-0.026	-0.091	-0.099	-0.025	-0.021	+0.021	-0.04	-0.067
rel. diff.	-27% (++)	-19% (++)	-17% (++)	-8% (+)	-5% 0	-12% (++)	-15% (++)	-24% (++)	-19% (++)	-9% (+)	-8% (+)	+7% (−)	-14% (++)	-41% (++)

Table A11 Nominal vector-field error $E_{F,i}$ (normalized, absolute) by state for the key interventions.

	v_x	v_y	r	ϕ	$\dot{\phi}$	θ_p	$\dot{\theta}_p$	z	$\dot{z}$	ω_{FL}	ω_{FR}	ω_{RL}	ω_{RR}	F_{w}
B: physics family vs agnostic														
comparator	0.269	0.393	2.7	0.00025	6.08	0.000243	14.7	0.000457	0.000821	8.92	4.65	5.81	4.51	0.000685
intervention	0.21	0.408	2.3	7.99e-05	6	0.000137	11	0.000401	0.000345	6.85	4.16	5.48	2.88	0.000811
abs. diff.	-0.059	+0.015	-0.4	-0.00017	-0.084	-0.00011	-3.8	-5.6e-05	-0.00048	-2.1	-0.5	-0.33	-1.6	+0.00013
rel. diff.	-21% ++	+5% (−)	-15% ++	-3% 0	-1% 0	+35% (−−)	-25% ++	+111% (−−)	+5% (−)	-23% ++	-10% (+)	-5% (+)	-35% ++	+89% (−−)
D: M1 Lagrangian														
comparator	0.229	0.42	3.02	0.000234	5.23	0.000381	13	0.00109	0.00076	6.98	3.91	7.19	2.54	0.000662
intervention	0.208	0.686	2.96	0.000588	10.7	0.000276	11	0.00105	0.000609	6.59	4.32	6.21	3.58	0.00157
abs. diff.	-0.021	+0.27	-0.061	+0.00035	+5.5	-0.0001	-2	-3.5e-05	-0.00015	-0.4	+0.41	-0.98	+1	+0.00091
rel. diff.	-8% (+)	+72% (—)	-0% 0	+3992% (−−)	+116% —	+1129% (—)	-15% ++	+102% (−−)	+533% (—)	-5% 0	+13% (—)	-13% ++	+43% —	+992% (—)
E: M2 primal–dual														
comparator	0.229	0.42	3.02	0.000234	5.23	0.000381	13	0.00109	0.00076	6.98	3.91	7.19	2.54	0.000662
intervention	0.234	0.505	2.33	9.42e-05	7.62	0.00013	12.1	0.000495	0.00134	6.07	3.21	5.89	3.77	0.00171
abs. diff.	+0.0044	+0.085	-0.68	-0.00014	+2.4	-0.00025	-0.93	-0.00059	+0.00058	-0.91	-0.7	-1.3	+1.2	+0.0011
rel. diff.	+5% 0	+26% (—)	-21% (++)	+2% 0	+53% (—)	+1807% (—)	-5% (+)	+29% (−−)	+213% (—)	-12% (++)	-15% (++)	-17% (++)	+50% —	+1320% (—)
F: LRA														
comparator	0.229	0.42	3.02	0.000234	5.23	0.000381	13	0.00109	0.00076	6.98	3.91	7.19	2.54	0.000662
intervention	0.211	0.336	2.98	0.000455	4.42	0.000124	11.9	0.00113	0.000907	6.79	3.86	6.71	2.83	0.000458
abs. diff.	-0.018	-0.084	-0.038	+0.00022	-0.8	-0.00026	-1.1	+4.5e-05	+0.00015	-0.19	-0.046	-0.48	+0.29	-0.0002
rel. diff.	-6% (+)	-16% (++)	+1% 0	+977% (−−)	-11% (++)	+1124% (—)	-6% (+)	+599% (−−)	+449% (—)	-1% 0	+3% 0	-6% (+)	+13% (—)	+171% (—)
G: Adam→L-BFGS-B (O1 median)														
comparator	0.227	0.396	2.68	0.000198	5.17	0.000157	12.5	0.000332	0.000346	7.05	4.02	6.69	2.8	0.000357
intervention	0.271	0.371	2.92	0.000151	5.11	0.000101	14.3	0.000226	0.000358	8.02	5.51	8.45	5.3	0.000231
abs. diff.	+0.043	-0.024	+0.24	-4.7e-05	-0.064	-5.6e-05	+1.8	-0.00011	+1.2e-05	+0.97	+1.5	+1.8	+2.5	-0.00013
rel. diff.	+19% —	-4% 0	+9% (−)	-10% (+)	+1% 0	-28% (++)	+15% —	-31% (++)	-14% (++)	+14% (—)	+38% —	+26% —	+91% —	-34% (++)
H: $l1$ prior, A0, 10%														
comparator	0.0584	0.0413	0.457	0.00024	0.506	0.000872	3.1	0.000239	0.00137	1.89	1.61	1.6	1.5	0.00459
intervention	0.0578	0.0287	0.459	1.07e-05	0.507	7.02e-05	3.19	0.000889	0.00291	1.89	1.68	1.57	1.51	0.00154
abs. diff.	-0.0006	-0.013	+0.0015	-0.00023	+0.0013	-0.0008	+0.093	+0.00065	+0.0015	-0.00021	+0.066	-0.031	+0.0075	-0.003
rel. diff.	-1% 0	-30% (++)	+1% 0	-95% (++)	+1% 0	+607% (—)	+3% 0	+958% (−−)	+284% (—)	-0% 0	+4% 0	-2% 0	+1% 0	-34% (++)
I: medium256, 10%														
comparator	0.0584	0.0413	0.457	0.00024	0.506	0.000872	3.1	0.000239	0.00137	1.89	1.61	1.6	1.5	0.00459
intervention	0.0186	0.0268	0.197	0.00116	0.317	0.000799	0.935	0.00208	0.0066	0.502	0.576	0.37	0.582	0.00291
abs. diff.	-0.04	-0.014	-0.26	+0.00092	-0.19	-7.3e-05	-2.2	+0.0018	+0.0052	-1.4	-1	-1.2	-0.92	-0.0017
rel. diff.	-68% (++)	-35% (++)	-57% (++)	+375% (—)	-37% (++)	+1909% (—)	-70% (++)	+1699% (—)	+7588% (—)	-73% (++)	-64% (++)	-77% (++)	-61% (++)	-37% (++)

Acknowledgments. This research was supported by the Software/Hardware Integration Lab (LISHA) at the Federal University of Santa Catarina (UFSC). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES).

References

[1] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019. https://doi.org/10.1016/j.jcp.2018.10.045

[2] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, pp. 422–440, 2021. https://doi.org/10.1038/s42254-021-00314-5

[3] S. Wang, Y. Teng, and P. Perdikaris, "Understanding and mitigating gradient flow pathologies in physics-informed neural networks," *SIAM Journal on Scientific Computing*, vol. 43, no. 5, pp. A3055–A3081, 2021. https://doi.org/10.1137/20M1318043

[4] S. Wang, X. Yu, and P. Perdikaris, "When and why PINNs fail to train: A neural tangent kernel perspective," *Journal of Computational Physics*, vol. 449, Art. no. 110768, 2022. https://doi.org/10.1016/j.jcp.2021.110768

[5] A. D. Jagtap, K. Kawaguchi, and G. E. Karniadakis, "Adaptive activation functions accelerate convergence in deep and physics-informed neural networks," *Journal of Computational Physics*, vol. 404, Art. no. 109136, 2020. https://doi.org/10.1016/j.jcp.2019.109136

[6] M. A. Nabian, R. J. Gladstone, and H. Meidani, "Efficient training of physics-informed neural networks via importance sampling," *Computer-Aided Civil and Infrastructure Engineering*, vol. 36, no. 8, pp. 962–977, 2021. https://doi.org/10.1111/mice.12685

[7] D. Wu, "Pool-based sequential active learning for regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1348–1359, 2019. https://doi.org/10.1109/TNNLS.2018.2868649

[8] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016. https://doi.org/10.1073/pnas.1517384113

[9] J. Schoukens and L. Ljung, "Nonlinear system identification: A user-oriented road map," *IEEE Control Systems Magazine*, vol. 39, no. 6, pp. 28–99, 2019. https://doi.org/10.1109/MCS.2019.2938121

[10] P. Tabuada and B. Gharesifard, "Universal approximation power of deep residual neural networks through the lens of control," *IEEE Transactions on Automatic Control*, vol. 68, no. 5, pp. 2715–2728, 2023. https://doi.org/10.1109/TAC.2022.3190051

[11] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation," *Journal of Machine Learning Research*, vol. 8, pp. 985–1005, 2007. https://www.jmlr.org/papers/v8/sugiyama07a.html

[12] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, Art. no. 91, 2006. https://doi.org/10.1186/1471-2105-7-91

[13] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010. https://www.jmlr.org/papers/v11/cawley10a.html

[14] L. F. O. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro, "Constrained learning with non-convex losses," *IEEE Transactions on Information Theory*, vol. 69, no. 3, pp. 1739–1760, 2023. https://doi.org/10.1109/TIT.2022.3187948

[15] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978. https://doi.org/10.1007/BF01588971

[16] J. Elenter, N. Naderializadeh, and A. Ribeiro, "A Lagrangian duality approach to active learning," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 37575–37589, 2022. https://doi.org/10.52202/068431-2724

[17] D. J. Rezende and F. Viola, "Generalized ELBO with constrained optimization, GECO," in *Third Workshop on Bayesian Deep Learning (NeurIPS)*, 2018. https://bayesiandeeplearning.org/2018/papers/33.pdf

[18] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018. https://proceedings.mlr.press/v80/chen18a.html

[19] A. A. Heydari, C. A. Thompson, and A. Mehmood, "SoftAdapt: Techniques for adaptive loss weighting of neural networks with multi-part loss functions," *arXiv preprint arXiv:1912.12355*, 2019. https://doi.org/10.48550/arXiv.1912.12355

[20] R. Bischof and M. A. Kraus, "Multi-objective loss balancing for physics-informed deep learning," *Computer Methods in Applied Mechanics and Engineering*, vol. 439, Art. no. 117914, 2025. https://doi.org/10.1016/j.cma.2025.117914

[21] P. Rathore, W. Lei, Z. Frangella, L. Lu, and M. Udell, "Challenges in training PINNs: A loss landscape perspective," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235 of *Proceedings of Machine Learning Research*, pp. 42159–42191, 2024. https://proceedings.mlr.press/v235/rathore24a.html