Physics-Aware Constraints for Inverse Shape Design

Anonymous authors
Paper under double-blind review

Abstract

Neural surrogates now carry computer aided engineering (CAE) into inverse shape design, where optimization descends the gradients of a frozen model. However, the optimizer and the judge are the same network: the search freely leaves the region where the surrogate has evidence, claiming gains that no data can verify. Today this trust gap is bridged by hand, with engineers specifying for every task where a shape may deform and how far it may drift. We present **Physics-Aware Constraints (PAC)**, the first supervision layer to read both constraints from the frozen GeoPT, pretrained on geometry lifted with synthetic dynamics. The core insight is that a model pretrained to perceive geometry through synthetic physical response should supervise the search rather than drive it. The model's weights bound how far optimization may drift off the family response manifold, and its distance field bounds where deformation is admissible, both calibrated on the family without labels. Across four CAE families, PAC consistently keeps optimization inside the trust band at little cost to legitimate gains, delivering on ship hulls a 64% wave drag reduction verified by an analytic oracle. The results separate the two sources of the constraint: its scale selectivity comes from the distance field representation and its detection sensitivity from physics lifted pretraining. PAC performs the judgment of a delivered design that autonomous design agents in industry still leave to an engineer, moving their full automation a step closer.

1 Introduction

Neural simulators are now efficient surrogates for numerical solvers in computer aided engineering (CAE), reducing expensive simulations to a single forward pass (Li et al., 2021; Pfaff et al., 2021; Wu et al., 2024; Adams et al., 2025). Industry has adopted them in commercial design software (Ansys, Inc., 2024; Altair Engineering Inc., 2024). The industrial payoff, however, lies in inverse shape design, the classical goal of aerodynamic optimization (Jameson, 1988), where a differentiable surrogate turns design into gradient descent over geometry (Remelli et al., 2020). Design agents now run this loop with little human input, proposing, evaluating and refining shapes on their own.

The central challenge is that the optimizer and the judge are the same network. Inverse design by backpropagation pushes the geometry wherever the predicted objective improves, the reward hacking of a learned objective (Amodei et al., 2016; Gao et al., 2023; Trabucco et al., 2021). We show that this surrogate hacking is more acute in CAE inverse design (Figure 1): an unconstrained optimizer reports over 50% improvement while drifting eleven times beyond the spread of the design family, and it concentrates almost all of the claimed gains on directions along which the training family never varied. Such gains cannot be verified, because no existing data can support or refute them.

The missing data cannot be generated on demand. Design loops increasingly run on published datasets rather than in house solvers (Ashton et al., 2024), and reproducing their numerical setup, from meshing to turbulence closure, is rarely possible outside the originating group. Inverse design in CAE is therefore offline model based optimization (Trabucco et al., 2021): the optimizer can never query the oracle. The only guarantee available is to confine optimization to the region where the surrogate carries evidence.

Existing safeguards fall into four families, none of which removes the manual burden. Conservative training penalizes the surrogate on out of distribution designs (Trabucco et al., 2021), at the price of retraining the production model. Trust regions size themselves from queries to the true objective

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

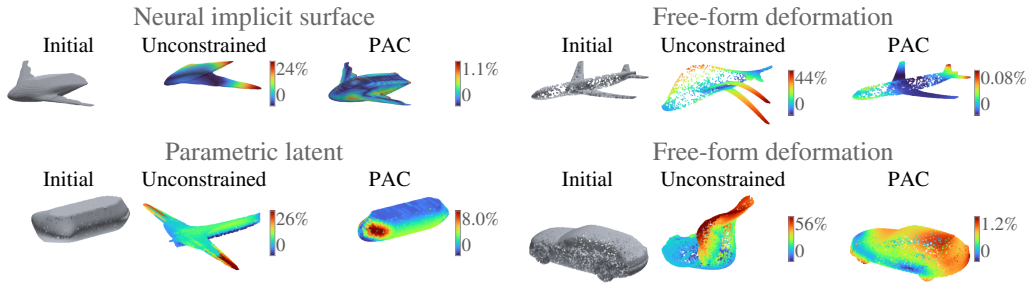


Figure 1: Inverse shape optimization with and without PAC on four CAE families. Columns are the initial design, the unconstrained optimum and the PAC optimum. Colour is the surface displacement from the initial, in per cent of the body diagonal.

(Eriksson et al., 2019), which an offline loop cannot make. Validity models filter abnormal geometries with a per task classifier (Li & Zhang, 2021), or keep latent space optimization near the data manifold through learned validity constraints, decoder uncertainty and diffusion priors (Griffiths & Hernández-Lobato, 2020; Notin et al., 2021; Chen et al., 2025), measuring geometric appearance or decodability rather than physical response. In industrial practice the burden falls back on engineers, who specify editable regions and deformation ranges by hand (Fan et al., 2026) and judge whether the delivered design can be trusted. That judgment is the one stage of the loop still performed by a person, and automating it needs a check that is physical rather than geometric and cheap enough to run inside the loop. What the engineer supplies answers two questions: how far a design may move before the surrogate should no longer be trusted, and where on the shape deformation is admissible at all.

We show that both answers are latent in the design family data, and that a single pretrained physics geometry model, GeoPT (Wu et al., 2026), provides the machinery to read them out. GeoPT is pretrained by lifting large scale geometry with synthetic dynamics, which couples its representation to physical response without physics labels, and it offers two readable layers. Its weights map any shape to a response fingerprint, so a deformation is charged by how it moves that response rather than by its geometric magnitude, and the whole design family is compressed into a compact set of response invariants. That response is the synthetic dynamics of pretraining and not the target physics the surrogate predicts, which is what makes it readable on any family without labels. The perception itself comes from the pretraining rather than the architecture, since a randomly initialized twin cannot reproduce it. Its distance field representation (Faugeras & Gomes, 2000; Park et al., 2019) places every member of the family in one metric frame, where the intrinsic variability of each surface location can be measured pointwise. On these two layers we build **Physics-Aware Constraints (PAC)**, a supervision layer that requires no labels and calibrates itself on held out members of the family. From the weights, a frozen differentiable constraint bounds how far optimization may drift from the family response manifold. From the representation, a pointwise deformability field bounds where deformation is admissible and by how much. PAC plugs into any differentiable inverse design loop, from free form deformation (Sederberg & Parry, 1986; Liu et al., 2026b) to neural implicit parametrizations (Remelli et al., 2020), and performs inside it the judgment that was left to the engineer. To our knowledge, this is the first time a pretrained physics geometry model enters the inverse design loop as a supervisor of gradients rather than a source of them. Our contributions are summarized as follows:

- We identify and quantify surrogate hacking in CAE inverse design, where unconstrained optimization drifts eleven times outside the region where the surrogate has evidence, leaving its gains unverifiable.
- We present PAC, a supervision layer that reads two trust constraints from one frozen pretrained physics geometry model, bounding how far a design may drift and where it may deform without labels.
- PAC keeps the search inside the trust band across four CAE design families and three parametrizations, delivering on the ship hull a design verified 64% lower in wave drag by

an analytic oracle, with factorial studies dissociating the two constraints and showing that only the pair ends the search at the band.

2 Related Work

Neural simulation for CAE. Neural surrogates have progressed from neural operators (Li et al., 2021; Liu et al., 2026a) and mesh based graph networks (Pfaff et al., 2021) to transformer solvers that scale to industrial geometries (Wu et al., 2024; Luo et al., 2025; Zhou et al., 2026; Adams et al., 2025). A parallel line makes the design side differentiable: adjoint methods established gradient based shape optimization (Jameson, 1988), free form deformation gave it a compact parametrization (Sederberg & Parry, 1986; Liu et al., 2026b), and differentiable iso surface extraction extended it to neural implicit shapes (Remelli et al., 2020). GeoPT extends this line by pretraining a transformer solver directly on distance field representations (Faugeras & Gomes, 2000; Park et al., 2019) lifted with synthetic dynamics (Wu et al., 2026). All of this machinery serves the forward direction. PAC gives the frozen pretrained model a second role on the inverse side, supervising the search under any differentiable parametrization.

Trustworthy inverse design. Offline model based optimization documents that gradient ascent on a learned score returns inputs with erroneously inflated predictions (Trabucco et al., 2021; Yu et al., 2021; Trabucco et al., 2022; Kim et al., 2026), but the family is defined on labelled designs over vector valued spaces with no spatial structure. Reward hacking is the same failure with a learned reward model (Skalse et al., 2022; Gao et al., 2023), and its standard remedy, penalizing deviation from a reference policy, is what the trust band supplies for shapes, with the family as the reference and a pretrained physics model as the metric. Two signals read from inside the loop do not transfer: a trust region sized by querying the true objective (Eriksson et al., 2019; Maus et al., 2022), and the surrogate's own error estimate (Lakshminarayanan et al., 2017; Ben Braiek et al., 2022), which asks the exploited model to police itself. Appendix F measures both. Validity is otherwise enforced per task, by classifiers over generated shapes (Li & Zhang, 2021), a subspace fitted to the family (Toal et al., 2010), decoder uncertainty or a learned manifold prior in latent space (Griffiths & Hernández-Lobato, 2020; Notin et al., 2021; Chen et al., 2025; Yang et al., 2025; Seiler et al., 2026), or hand specified editable regions (Hajdik et al., 2023; Fan et al., 2026; Sun et al., 2026): each bounds how far in one learned space, or answers where by hand. PAC reads both from one frozen pretrained model, without labels and without retraining the production surrogate.

3 Problem Setup

We consider inverse shape design within one engineering family, a set $\mathcal{G}$ of shapes sharing a common structure such as blended wing bodies, observed as a finite sample $\{G_i\}_{i=1}^N$ without physics labels, the only data PAC uses. Under an operating condition $S \in \mathcal{S}$ that drives the system, a numerical solver produces the field $\mathbf{u}(x; G, S) \in \mathbb{R}^{C_u}$ at surface points x, from which a differentiable functional J reads a scalar design objective. The condition S is held fixed while the shape varies, the steady state regime of industrial design. In aerodynamics, for example, G is a wing surface, S the Mach number and incidence, $\mathbf{u}$ the surface pressure and shear, and J their integral along the flow direction.

A forward surrogate $f_\phi(G, S)$, trained on the family and frozen throughout, estimates $\mathbf{u}$ directly. Inverse design optimizes a parametrization $\theta \in \mathbb{R}^d$ mapped to geometry by a differentiable $G(\cdot) : \mathbb{R}^d \to \mathcal{G}$, such as free form deformation displacements or the latent code of a neural implicit surface, initialized at $G_0 \in \mathcal{G}$. Since both are differentiable, one minimizes the predicted objective by gradient descent,

$$\theta^\star = \arg\min_\theta \; J\big(f_\phi(G(\theta), S)\big). \tag{1}$$

Nothing in Eq. 1 constrains θ, while f_ϕ estimates $\mathbf{u}$ reliably only near the family it was trained on. PAC (Section 4) adds two constraints read from a frozen pretrained physics geometry model, separate from f_ϕ, and calibrated on $\{G_i\}$ without labels, confining the search to the region where f_ϕ carries evidence. Trust is meant in this operational sense throughout, a design being trusted when it lies inside a band calibrated on the family, and whether the surrogate is also more accurate there is measured separately (Appendix F.1).

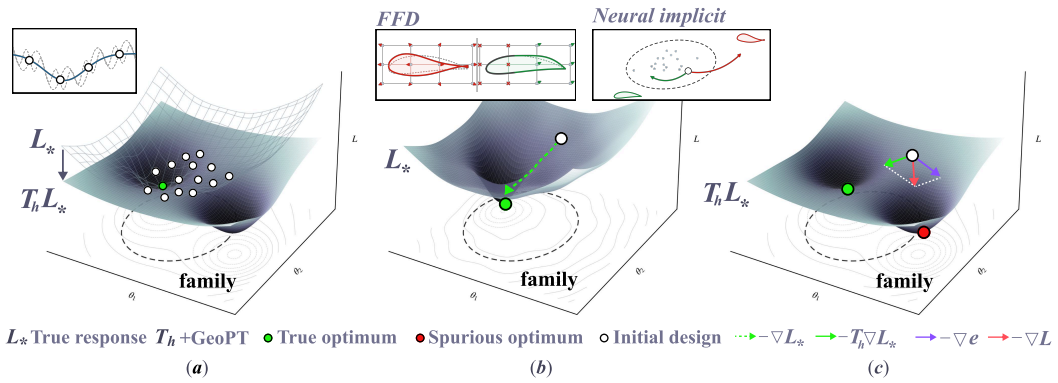


Figure 2: One response landscape $L = T_h L_\star + e$, read twice. (a) Forward: pretraining smooths $L_\star$, which few labels then fix. (b) Inverse: the gradient is read on the sharp $L_\star$. (c) On the surrogate the residual $-\nabla e$ overwhelms the flattened signal and $-\nabla L$ leaves the family.

4 Method

PAC reads a trust boundary from the frozen model's weights and a deformability field from its representation, both calibrated on the family, and holds the search of Eq. 1 inside them. Figure 2 shows the landscape this rests on and Figure 3 where the two constraints attach in the design loop.

4.1 From Gradient to Constraint

Whether the response of GeoPT should steer the search turns on how it enters the loop, which a single modeling assumption settles. Since GeoPT pretrains on over a million shapes by predicting geometric feature trajectories that vary smoothly across them (Wu et al., 2026), we model its effect as viewing the true objective landscape $L_\star(\theta) = J\big(\mathbf{u}(G(\theta), S)\big)$ through a smoothing operator,

$$L(\theta) \;=\; (T_h L_\star)(\theta) \;+\; e(\theta), \tag{2}$$

where T_h acts on functions of θ at a scale h that pretraining increases, and e is the finetuning residual, with $\sigma^2 = \mathbb{E}\|\nabla_\theta e\|^2$ the mean square of its gradient.

Value and gradient. Forward prediction reads the value $T_h L_\star$, and smoothing trades bias for lower variance, so in the low label regime a larger h lowers the dominant term. Inverse design instead descends the gradient $\nabla_\theta L = T_h \nabla_\theta L_\star + \nabla_\theta e$ and succeeds not on the value of L but on the fidelity of its direction to the true one, $\rho(\theta) = \langle \nabla_\theta L, \nabla_\theta L_\star \rangle / (\|\nabla_\theta L\| \|\nabla_\theta L_\star\|)$, which must be positive to make progress at all.

The forward-inverse conflict. The gradient splits into a signal $T_h \nabla_\theta L_\star$ of scale $s = \|T_h \nabla_\theta L_\star\|$, the response sharpness, and a noise $\nabla_\theta e$ of scale σ, set by the residual. Treating the residual gradient as isotropic and uncorrelated with the signal, and writing $\cos\psi$ for the alignment of the smoothed gradient $T_h \nabla_\theta L_\star$ with the true one,

$$\rho \;\approx\; \frac{s\,\cos\psi}{\sqrt{s^2 + \sigma^2}}, \tag{3}$$

which increases with s at fixed alignment, $\cos\psi \le 1$ only lowering it further. Raising h lowers s and leaves σ untouched: by the forward paragraph this improves generalization, by Eq. 3 it lowers ρ, and descending the blunted gradient slides the design off the family (Figure 2). Section 5.2 measures both halves of this conflict on the same checkpoints.

Constrained formulation. Because the gradient of $T_h L_\star$ is unreliable while its value is not, we constrain the search with the value rather than descend the gradient. PAC replaces the unconstrained Eq. 1 with

$$\theta^\star \;=\; \arg\min_\theta \; J\big(f_\phi(G(\theta), S)\big) \quad \text{s.t.} \quad c\big(\Phi(G(\theta))\big) \le \tau, \tag{4}$$

where Φ maps a shape to the frozen model's reading of it at a fixed probe set (Section 4.2) and c scores that reading against the family.

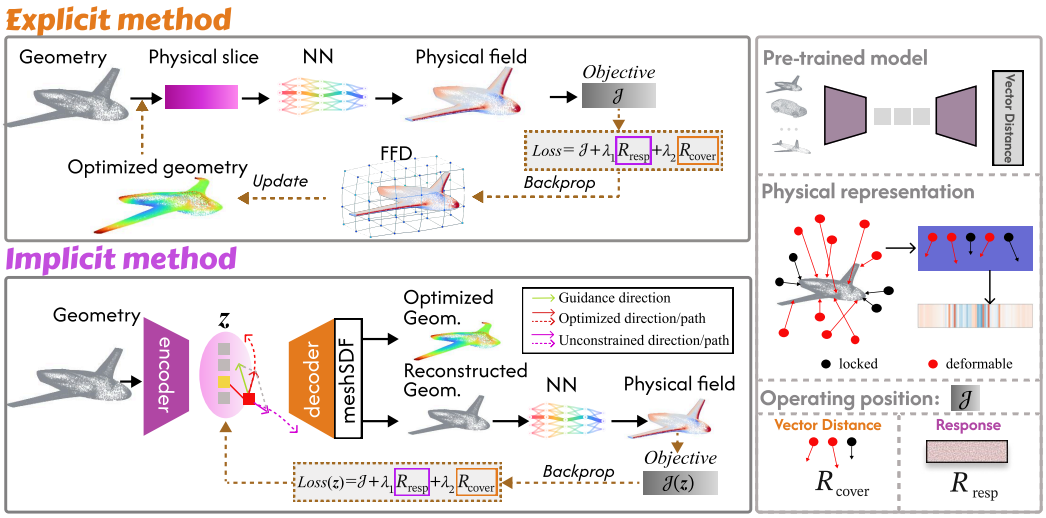


Figure 3: Where the two constraints enter the design loop. Top, the explicit path, free form deformation. Bottom, the implicit path, a neural implicit surface. Right, the frozen pretrained model and the two readouts taken from it: the vector distance field, black where the family holds the surface still and red where it varies it, and the response fingerprint. Both paths carry them in the loss as $\lambda_1 R_{\text{resp}}$ and $\lambda_2 R_{\text{cover}}$.

4.2 Two Trust Constraints

The vector distance field. Both constraints perceive the shape through the vector distance field, which at a query point x returns the vector to the nearest surface point of G, $v_G(x) = p^*_G(x) - x$ with $p^*_G(x) = \arg\min_{p \in \partial G} \|p - x\|$, a directional and differentiable encoding of geometry (Faugeras & Gomes, 2000). GeoPT is pretrained to predict this field, so stacking its frozen readout over a fixed probe set gives the fingerprint $\Phi(G)$ of Eq. 4. Where pretraining treats the field as a target, we take it as the medium in which trust is measured: it is boundary driven, so a deformation the family cares about registers in v_G, and in the far field it is governed by overall scale, so it records the proportions a family shares.

Response manifold constraint. The fingerprints of the family occupy a manifold. Let $d(\theta)$ be the distance from $\Phi(G(\theta))$ to it, estimated as the mean distance to its five nearest family members. We set the trust boundary τ to the 95th percentile of d over held out family shapes, a conformal quantile whose coverage Appendix A.1 bounds, and penalize only excursions past it,

$$R_{\text{resp}}(\theta) \;=\; \text{relu}\big(d(\theta)/\tau - 1\big)^2. \tag{5}$$

The gradient of R_{resp} is the only one taken through the frozen model, and it only restores the design toward the family, never supplying the descent direction of J. Appendix C.2 gives the calibration that keeps d comparable across shapes.

Deformability field. Where a shape may deform is a property of the family, not of any one member. The spread of the vector distance field across it, $w(x) = \big(\mathbb{E}_G \|v_G(x) - \bar{v}(x)\|^2\big)^{1/2}$ with $\bar{v}(x) = \mathbb{E}_G\, v_G(x)$, measures how much the family varies at each surface point, with w_0 the noise floor of the estimate. Small w marks regions the family holds fixed, a wing's leading edge or an engine mount, large w regions it varies freely.

Masks and budgets. The field becomes a per-point mask $m(x) = \text{clip}\big((w(x) - w_0)/((k-1)\,w_0),\, 0,\, 1\big)$, which locks a point where the family holds it still ($w \le w_0$) and frees it where the family clearly moves it ($w \ge k\, w_0$). The mask multiplies the displacement, and an amplitude budget b is read from the same statistic,

$$\tilde{\delta}(x) \;=\; m(x)\,\delta(x), \qquad \max_x \|\tilde{\delta}(x)\| \le b, \qquad b = \kappa\, q_{90}(w), \tag{6}$$

so the search moves each point in proportion to how much the family moves it and is capped in proportion to how much the family varies rather than by a fixed fraction of the geometry. The boundary τ, the mask m and the budget b are all read from the held out family. The only constants are the two relative multipliers $k = \kappa = 1.5$, fixed once for all families. The optimizer settings are listed in Appendix C.2.

4.3 Trust-Constrained Optimization

Unified objective. The two constraints enter the loop as two penalties: the response manifold as R_{resp} (Eq. 5), and the deformability field as a coverage penalty R_{cover} that keeps in place the regions the family holds fixed. With the displacement masked and budgeted as in Eq. 6, PAC minimizes

$$\mathcal{L}(\theta) \;=\; J\big(f_\phi(G(\theta), S)\big) \;+\; \lambda_1 |J_0|\, R_{\text{resp}}(\theta) \;+\; \lambda_2 R_{\text{cover}}(\theta) \;+\; \gamma\, \Omega(\theta), \tag{7}$$

where Ω is a smoothness regularizer and J_0 the objective at the initial design. Scaling by $|J_0|$ makes λ_1 dimensionless, so one value transfers across families whose objectives differ by orders of magnitude (Section 5). Where the parametrization admits no gradient, as for the ship hull design vector, the same two readouts act as a gate on each step and as a filter on candidates instead of as penalties.

Attachment. Where each constraint attaches follows the parametrization. Under free form deformation (Sederberg & Parry, 1986) the mask acts at the displacement itself, the hardest form available: it zeroes locked regions, so R_{cover} vanishes and no gradient reaches a control point from a locked region. Under a neural implicit surface (Park et al., 2019; Remelli et al., 2020) the code is global and admits no such zeroing. Surface points move as $\dot{p} = -\nabla f\,(\partial_\theta f)/\|\nabla f\|^2$, and the deformability field enters as a one way Chamfer anchor

$$R_{\text{cover}}(\theta) \;=\; \sum_{x \in \mathcal{X}_0} \min_{p \in G(\theta)} \big\|x - p\big\|^2, \tag{8}$$

requiring the current surface to keep covering the region $\mathcal{X}_0$ that the family holds fixed. Appendix B details each path.

Delivery protocol. We deliver the best feasible iterate rather than the last, of smallest J among those with $R_{\text{resp}} = 0$ and admitted by the parametrization, with the step size annealed.

5 Experiments

We evaluate PAC on four CAE design families spanning three parametrizations, one of which carries an analytic ground truth. Throughout, **bold** marks the best value in a column and underline the second.

Families and parametrizations. The blended wing body (BWB) is optimized through a neural implicit surface (Park et al., 2019; Remelli et al., 2020), the NASA common research model (CRM) and the car through a free form deformation (FFD) lattice (Sederberg & Parry, 1986), and the ship hull through its 45 dimensional design vector. Each family carries its own frozen surrogate. The objective is a differentiable surface integral of the predicted field rather than a scalar coefficient head. For the ship hull an analytic wave drag model (Michell, 1898) and naval feasibility constraints act as an oracle the optimizer never sees.

Trust baselines. We compare four ways of turning a shape into the vector the constraint measures. A voxel occupancy descriptor isolates the calibration protocol, a randomly initialized twin of GeoPT isolates pretraining, and a ShapeVAE encoder pretrained for generation on large scale geometry (Tencent Hunyuan3D Team, 2025), run with its own twin, isolates geometry only pretraining. All four share the same reference shapes, standardization, distance, trust boundary and deformation seeds, each under a configuration fixed before the comparison.

Claimed, trusted, verified. We report three levels of evidence. A *claimed* improvement is what the surrogate reports at an arbitrary design. A *trusted* improvement is the claimed improvement of the best iterate satisfying $R_{\text{resp}} = 0$. A *verified* improvement is measured by the analytic oracle. Intervals are bootstrapped over 3000 resamples of the 80 held out geometries, or paired over family members for coverage. Appendix C.2 gives full configurations.

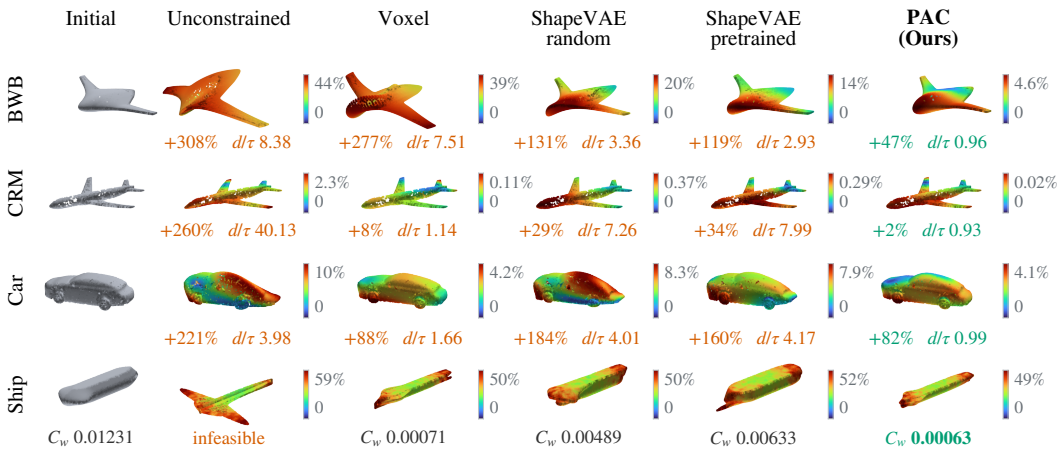


Figure 4: Designs delivered by the unconstrained arm and four trust backends. The ship row filters a generated population instead. Colour is the surface displacement from the initial, in per cent of the body diagonal, and d/τ is read against one common trust boundary. See Appendix D.

5.1 Main Results

Surrogate hacking. Table 1 shows the same failure mode on all four families.

(i) Inflated claims off the band. Unconstrained optimization reports large improvements far outside the region where the surrogate has evidence: $+55.7\%$ at 11.7 times the trust boundary on the blended wing body, $+1290.7\%$ at 639 times on CRM, $+526.8\%$ at 12.7 times on the car, and a vanishing wave drag on the ship hull that the oracle rules infeasible. Changing only the seed moves these claims by up to a third of their own value.

(ii) Gains on never varied directions. The deformations behind them move the shape where the family never varies: the unconstrained arm breaks the mirror symmetry every blended wing body obeys and collapses the ship hull into a thin slab, cutting draught by 64%. No data could support or refute such a change. PAC instead delivers inside the band on all four families.

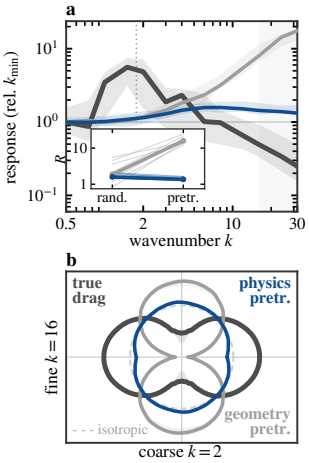


Figure 5: Induced metric of each trust backend: (a) by ripple wavenumber, (b) by deformation direction.

Backend comparison. Figure 4 holds the loop fixed and changes only the backend that measures trust. Unconstrained, the search claims up to $+308\%$, ends as far as forty times the boundary, and thickens the blended wing body by 77% along its thinnest axis. The generative encoder delivers outside the band whether or not it is pretrained, at 2.9 to 8.0 times the boundary, while ours delivers inside it on all three rows: the backend, not the loop, decides whether a delivered design can be trusted. Figure 5 shows the metric behind this: the generative encoder still responds 16 times where the true wave drag has lost 90% of its sensitivity.

Trust regions. On the design space plane of Figure 6, every fence rejects the unconstrained delivery, at 1.5 to 5.9 times its boundary. PAC keeps the most family inside its band, 58% against 52% for the voxel descriptor and 36–38% for the generative encoder, whose induced metric charges the family's own variation as an excursion, and all paired differences exclude zero. On the blended wing body plane of Figure 13 it holds 88% of the members in a region 3.7 times smaller than the voxel band and rejects the exploit the deepest.

Factorial dissociation. Removing each constraint in turn on the neural implicit path, where both apply, dissociates them (Table 2).

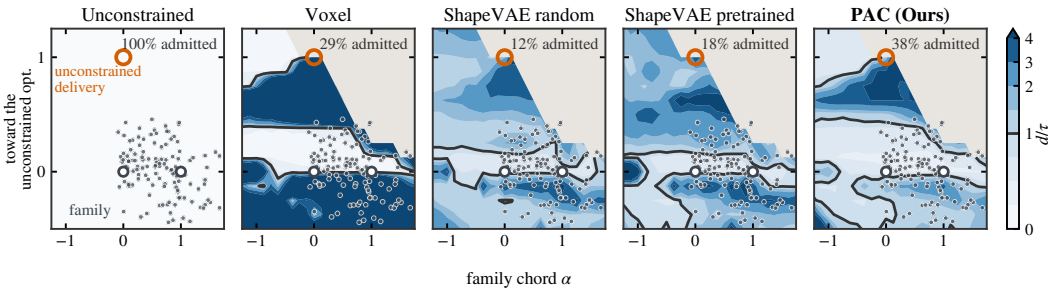


Figure 6: Trust regions on one plane of the ship hull design space, spanned by a chord between two family members and by the blend toward the unconstrained delivery (ring). Colour is d/τ under each backend's own calibration. Dots are family members and grey is geometry the parametrization cannot produce. Annotations give the share of the plane admitted.

Table 1: Inverse optimization, d/τ at the reported design. The ship hull row is oracle verified (full version: Table 7).

	Unconstrained		**PAC (Ours)**	
Family	claimed	d/τ	trusted↑	d/τ
BWB	+55.7%	11.74	+5.1%	0.86
CRM	+1290.7%	639	+2.2%	0.85
Car	+526.8%	12.69	+22.7%	0.70
Ship hull	infeasible	1.64	+64.5%	0.47

Table 2: Factorial dissociation of the two constraints on the neural implicit path. Trusted is undefined without the fence (full version: Table 8).

Constraint		Search outcome		
fence	field	trusted↑	terminal d/τ ↓	in band↑
—	—	/	11.74	2
✓	—	+8.8%	1.10	6
—	✓	+5.1%	1.62	16
✓	✓	+5.1%	**0.97**	**24**

(i) The fence bounds how far, not where. With the response manifold constraint alone the best in band iterate delivers 8.8%, but the search still ends outside the band at 1.10 and disturbs the regions the family holds fixed as much as the free ones.

(ii) The field pins where, not how far. With the deformability field alone those regions are protected, yet the search ends at 1.62 times the boundary. Being pointwise, the field intercepts the local edits the fence is structurally blind to, cutting locked region displacement to 0.310 of the free regions under free form deformation (Appendix D).

(iii) Only the pair ends at the band. The combination terminates inside the band, at 0.97, spending 24 of 36 iterates in band against 16, 6 and 2. Over the seven runs of Appendix D it ends at a median of 1.11 against 1.62 and 1.72 for either constraint alone, inside the band in three where neither alone ever does, and delivers the field arm's iterate in six.

Oracle verification. On the ship hull the analytic evaluator settles what the surrogate cannot. The surrogate's descent leaves the naval feasible set within a few steps and claims a vanishing wave drag at a hull the oracle rules infeasible, whereas the hull PAC delivers, the surrogate's best iterate among those the parametrization admits and the fence keeps in band, is feasible with a wave drag 64.5% below the initial hull (Table 1, Appendix F.3). Used as a label free filter on a generated population (Figure 7), the same constraint keeps candidates whose median true wave drag is 17.4 times lower than the population median at the four kept, and its first choice carries the lowest true drag of any backend, 2.5 times below the next best.

5.2 Constraint Analysis

Inverse gradient fidelity. The premise of Section 4.1, illustrated in Figure 2, is that the pretrained model cannot supply the descent direction. Over 64 deformation directions the gradient of a surrogate finetuned from GeoPT attains a rank correlation of +0.11 with the true descent against +0.27 for the same architecture trained from scratch, while the same checkpoints cut few label prediction error by up to 90%. Pretraining compresses the early layer response to a 2% deformation by 5.6 times and flattens the landscape by 2.2 times: one operator, one scale, opposite signs (Eq. 3).

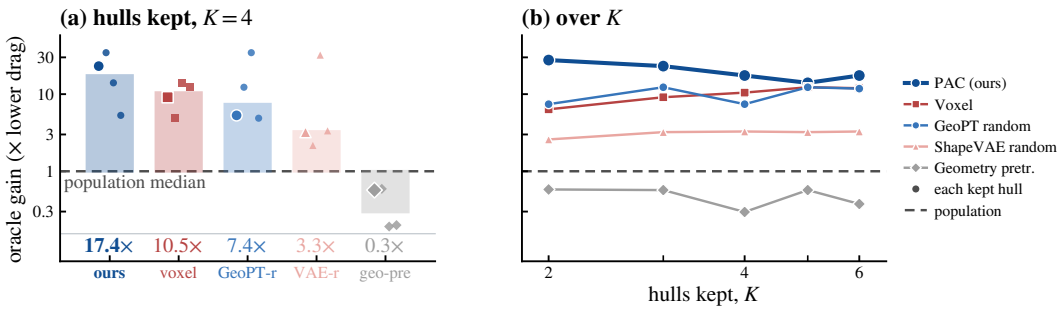


Figure 7: Label free selection on the ship hull, scored by the analytic oracle. Each backend keeps the K hulls it ranks closest to the family, and the oracle reads their true wave drag, reported as the gain over the population median. (a) $K = 4$, one point per kept hull, the first ranked drawn larger. (b) Gain against K, out to 6 of 115 candidates.

Table 3: Composition of the trust backend. Detection AUC separates a 2% deformation from the clean shape. Band consumed is the share of the trust boundary moved by a harmless input change, median over held out shapes. Oracle gain is that of Figure 7 at $K = 4$. Δ is pretrained minus its random twin.

	Detection AUC ↑				Band consumed	Oracle
Backend	FFD 4^3	FFD 8^3	Bump	Ripple	(%) ↓	gain ↑
Voxel	0.79	0.82	0.78	0.89	40.3	10.5×
GeoPT, random	0.68	0.71	0.63	0.91	15.6	7.4×
GeoPT, pretrained (Ours)	**0.92**	**0.95**	0.73	0.96	19.0	**17.4**×
ShapeVAE, random	0.91	0.93	**0.79**	0.87	**6.3**	3.3×
ShapeVAE, pretrained	0.90	0.93	0.73	**1.00**	228.0‡	0.3×
Δ GeoPT	**+0.245**	**+0.241**	**+0.095**	+0.054	**+3.4**	**+10.0**×
Δ ShapeVAE	−0.009	+0.001	−0.057	**+0.126**	+221.7	−3.0×

‡ The harmless change alone exceeds the whole band, on every one of the held out shapes.

Effect of pretraining. Table 3 separates the protocol from pretraining by comparing each pretrained backend with its own randomly initialized twin. Physics lifted pretraining raises detection on all four deformation bands, by $+0.054$ to $+0.245$ with intervals excluding zero, where geometry only pretraining raises one band of four. The scale selectivity of the fingerprint comes with the distance field representation, which charges fine and coarse ripples alike (Figure 5a inset), where geometry only pretraining creates the fine scale overcharge (1.9 to 15.9). Pretraining adds detection on top of that selectivity (Appendix E). The backends also differ in what a harmless change costs: a 0.3% point cloud noise crosses the whole trust band on every held out shape for the generative encoder and on 39% for the voxel descriptor, against none for GeoPT (Appendix F.4).

Self-calibration. Substituting a fixed fraction of the geometry for the family read budget of Eq. 6 removes every feasible iterate on CRM (Appendix C.2).

6 Conclusion

Physics-Aware Constraints keep inverse shape design trustworthy where the oracle cannot be queried. Read without labels from one frozen pretrained physics geometry model, the two constraints bound how far a design may drift from the family response manifold and where it may deform. Across four CAE design families they hold the search inside the calibrated band, factorial studies dissociate them, and an analytic oracle certifies the trusted optimum as the hull of lower true wave resistance. Because both are read from the family rather than specified per task, the loop transfers across families and parametrizations with no human in it.

Ethics statement

We adhere to the ICLR Code of Ethics. This work involves no human participants, personally identifiable information, or sensitive content.

AI use statement

We used a large language model (LLM) for language polishing, for example by correcting grammar and refining phrasing for greater clarity and precision. The LLM also assisted in organizing and adjusting parts of the code. All edits suggested by the LLM were reviewed and approved by the authors, and we take responsibility for the final content of this work, including any text or code produced with the aid of the LLM.

Reproducibility statement

The constraint formulations and algorithms are detailed in Appendix B, with experimental configurations in Appendix C. Full results and repeated runs are reported in Appendix D.

References

Corey Adams, Rishikesh Ranade, Ram Cherukuri, and Sanjay Choudhry. GeoTransolver: Learning physics on irregular domains using multi-scale geometry aware physics attention transformer. *arXiv preprint arXiv:2512.20399*, 2025.

Altair Engineering Inc. Altair PhysicsAI. `https://www.siemens.com/en-us/products/simcenter/engineering-data-science-ai/physicsai/`, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Ansys, Inc. Ansys SimAI. `https://ansys.synopsys.com/products/ai/simai`, 2024.

Neil Ashton, Charles Mockett, Marian Fuchs, Louis Fliessbach, Hendrik Hetmann, Thilo Knacke, Norbert Schonwald, Vangelis Skaperdas, Grigoris Fotiadis, Astrid Walle, Burkhard Hupertz, and Danielle Maddix. DrivAerML: High-fidelity computational fluid dynamics dataset for road-car external aerodynamics. *arXiv preprint arXiv:2408.11969*, 2024.

Noah J. Bagazinski and Faez Ahmed. Ship-D: Ship hull dataset for design optimization using machine learning. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2023.

Houssem Ben Braiek, Ali Tfaily, Foutse Khomh, Thomas Reid, and Ciro Guida. SmOOD: Smoothness-based out-of-distribution detection approach for surrogate neural networks in aircraft design. In *IEEE/ACM International Conference on Automated Software Engineering*, 2022.

Giovanni Catalani, Jean Fesquet, Xavier Bertrand, Frédéric Tost, Michael Bauerheim, and Joseph Morlier. Towards scalable surrogate models based on neural fields for large scale aerodynamic simulations. *arXiv preprint arXiv:2505.14704*, 2025.

Long Chen, Emre Oezkaya, Jan Rottmayer, Nicolas R. Gauger, Zebang Shen, and Yinyu Ye. Adjoint-based aerodynamic shape optimization with a manifold constraint learned by diffusion models. *arXiv preprint arXiv:2507.23443*, 2025.

Mohamed Elrefaie, Florin Morar, Angela Dai, and Faez Ahmed. DrivAerNet++: A large-scale multimodal car dataset with computational fluid dynamics simulations and deep learning benchmarks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

David Eriksson, Michael Pearce, Jacob R. Gardner, Ryan D. Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems*, 2019.

Wenhao Fan, Yuanwei Bin, Jianghan Gu, Wenfa Luo, Jiao Xiang, Yuntian Chen, and Shiyi Chen. Knowledge-constrained shape optimization with a mixture-of-experts neural operator for high-confidence design. *arXiv preprint arXiv:2607.09763*, 2026.

Olivier Faugeras and José Gomes. Dynamic shapes of arbitrary dimension: The vector distance functions. In *The Mathematics of Surfaces IX*. Springer, 2000.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, 2023.

Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chemical Science*, 11(2):577–586, 2020.

Hannah M. Hajdik, Anil Yildirim, Ella Wu, Benjamin J. Brelje, Sabet Seraj, Marco Mangano, Joshua L. Anibal, Eirikur Jonsson, Eytan J. Adler, Charles A. Mader, Gaetan K. W. Kenway, and Joaquim R. R. A. Martins. pyGeo: A geometry package for multidisciplinary design optimization. *Journal of Open Source Software*, 8(87):5319, 2023.

Antony Jameson. Aerodynamic design via control theory. *Journal of Scientific Computing*, 3(3): 233–260, 1988.

Minsu Kim, Jiayao Gu, Ye Yuan, Taeyoung Yun, Zixuan Liu, Yoshua Bengio, and Can Chen. Offline model-based optimization: Comprehensive review. *Transactions on Machine Learning Research*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Jichao Li and Mengqi Zhang. On deep-learning-based geometric filtering in aerodynamic shape optimization. *Aerospace Science and Technology*, 112:106603, 2021.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Shouyi Liu, Xiaokang Yang, and Yuntian Chen. Riesz neural operator for solving partial differential equations. In *International Conference on Learning Representations*, 2026a.

Ye Liu, Shouyi Liu, Ding Wang, Huiyu Yang, Ruizhe Deng, Qian Li, Yuxiao Hu, Jianghang Gu, Yongzheng Liu, Quanshi Zhang, Shiyi Chen, and Yuntian Chen. UniDrag: Unified multi-field prediction and robust shape optimization for vehicle aerodynamics. In *International Conference on Machine Learning*, 2026b.

William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Proceedings of SIGGRAPH*, pp. 163–169, 1987.

Huakun Luo, Haixu Wu, Hang Zhou, Lanxiang Xing, Yichen Di, Jianmin Wang, and Mingsheng Long. Transolver++: An accurate neural solver for PDEs on million-scale geometries. In *International Conference on Machine Learning*, 2025.

Natalie Maus, Haydn T. Jones, Juston S. Moore, Matt J. Kusner, John Bradshaw, and Jacob R. Gardner. Local latent space Bayesian optimization over structured inputs. In *Advances in Neural Information Processing Systems*, 2022.

John Henry Michell. The wave-resistance of a ship. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 45(272):106–123, 1898.

Pascal Notin, José Miguel Hernández-Lobato, and Yarin Gal. Improving black-box optimization in VAE latent space using decoder uncertainty. In *Advances in Neural Information Processing Systems*, 2021.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Edoardo Remelli, Artem Lukoianov, Stephan R. Richter, Benoît Guillard, Timur Bagautdinov, Pierre Baque, and Pascal Fua. MeshSDF: Differentiable iso-surface extraction. In *Advances in Neural Information Processing Systems*, 2020.

Thomas W. Sederberg and Scott R. Parry. Free-form deformation of solid geometric models. In *Proceedings of SIGGRAPH*, pp. 151–160, 1986.

Emilien Seiler, Nicolas Talabot, Yingxuan You, Federico Stella, and Pascal Fua. Flow-corrected shape optimization: Taming manifold drift in high-dimensional 3D models. *arXiv preprint arXiv:2608.07199*, 2026.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, 2022.

Philippe R. Spalart and Steven R. Allmaras. A one-equation turbulence model for aerodynamic flows. In *30th AIAA Aerospace Sciences Meeting and Exhibit*, 1992.

Guoze Sun, Tianya Miao, Haoyang Huang, Huaguan Chen, Han Wan, Rui Zhang, and Hao Sun. Geometry-aware neural optimizer for shape optimization and inversion. In *International Conference on Machine Learning*, 2026.

Nicholas Sung, Steven Spreizer, Mohamed Elrefaie, Kaira Samuel, Matthew C. Jones, and Faez Ahmed. BlendedNet: A blended wing body aircraft dataset and surrogate model for aerodynamic predictions. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2025.

Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

David J. J. Toal, Neil W. Bressloff, Andy J. Keane, and Carren M. E. Holden. Geometric filtration using proper orthogonal decomposition for aerodynamic design optimization. *AIAA Journal*, 48 (5):916–928, 2010.

Brandon Trabucco, Aviral Kumar, Xinyang Geng, and Sergey Levine. Conservative objective models for effective offline model-based optimization. In *International Conference on Machine Learning*, 2021.

Brandon Trabucco, Xinyang Geng, Aviral Kumar, and Sergey Levine. Design-bench: Benchmarks for data-driven offline model-based optimization. In *International Conference on Machine Learning*, 2022.

John Vassberg, Mark DeHaan, Melissa Rivers, and Richard Wahls. Development of a common research model for applied CFD validation studies. In *26th AIAA Applied Aerodynamics Conference*, 2008. AIAA Paper 2008-6919.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *International Conference on Machine Learning*, 2024.

Haixu Wu, Minghao Guo, Zongyi Li, Zhiyang Dou, Mingsheng Long, Kaiming He, and Wojciech Matusik. GeoPT: Scaling physics simulation via lifted geometric pre-training. In *International Conference on Machine Learning*, 2026.

Haoyu Yang, Kamyar Azizzadenesheli, and Haoxing Ren. Deep generative prior for first order inverse optimization. *arXiv preprint arXiv:2504.20278*, 2025.

Sihyun Yu, Sungsoo Ahn, Le Song, and Jinwoo Shin. RoMA: Robust model adaptation for offline model-based optimization. In *Advances in Neural Information Processing Systems*, 2021.

Hang Zhou, Haixu Wu, Haonan Shangguan, Yuezhou Ma, Huikun Weng, Jianmin Wang, and Mingsheng Long. Transolver-3: Scaling up transformer solvers to industrial-scale geometries. In *International Conference on Machine Learning*, 2026.

A Derivations

This appendix derives the two properties the main text uses without proof: the coverage guarantee behind the trust boundary τ (Section 4.2) and the level set sensitivity behind the neural implicit path (Section 4.3).

A.1 Coverage of the Trust Band

The boundary τ is read from held out family members, and the guarantee it carries is distribution free.

Proposition 1. *Let $d_1, \dots, d_n$ be the fingerprint distances of n held out family members and d_{n+1} that of a fresh member, the $n{+}1$ values exchangeable. Set $\tau = d_{(k)}$, the k-th smallest calibration distance, with $k = \lceil (1-\alpha)(n+1) \rceil$. Then*

$$\Pr\left[\, d_{n+1} \le \tau \,\right] \;\ge\; 1 - \alpha. \tag{9}$$

Proof. Exchangeability makes the rank of d_{n+1} among the $n{+}1$ values uniform on $\{1, \dots, n{+}1\}$, ties broken at random. The event $d_{n+1} \le d_{(k)}$ contains the event that this rank is at most k, so its probability is at least $k/(n+1) \ge 1 - \alpha$. □

This is the conformal calibration guarantee (Vovk et al., 2005) applied to the fingerprint distance. It delimits exactly what τ certifies. A genuine family member is rejected by the fence with probability at most α, with no assumption on the distribution of d and no labels. At the held out sizes of Table 5 the bound $k/(n+1)$ is $39/41 = 0.951$ for the largest sets and $7/8 = 0.875$ for the smallest, whose size caps the achievable level at $k = n$. The empirical percentile used in practice lies within one order statistic of $d_{(k)}$. Power off the family, and whether surrogate error is in fact lower inside the band, are separate empirical questions, measured in Table 3 and Appendix F.1.

A.2 Level Set Sensitivity

The neural implicit path needs the first order motion of a surface point as the code changes. Let $f(x, \theta)$ be the decoded signed distance field and $p(\theta)$ a point on its zero level set, $f(p(\theta), \theta) = 0$. Differentiating along θ,

$$\nabla_x f \cdot \dot{p} \;+\; \partial_\theta f \;=\; 0, \tag{10}$$

which constrains only the normal component of $\dot{p}$, since motion tangent to the level set changes no shape. The minimum norm solution is

$$\dot{p} \;=\; -\,\frac{\nabla_x f}{\|\nabla_x f\|^2}\,\partial_\theta f, \tag{11}$$

the sensitivity used in Section 4.3. The MeshSDF sensitivity $\dot{p} = -\nabla_x f\, \partial_\theta f$ (Remelli et al., 2020) is the special case $\|\nabla_x f\| = 1$, exact for an ideal signed distance function. A learned decoder satisfies the eikonal property only approximately. Wherever $\|\nabla_x f\|$ deviates from one, the uncorrected formula scales the true motion by $\|\nabla_x f\|^2$. It inflates the descent where the decoder is steep, and applied to the fence gradient it weakens the constraint exactly where it fires. Eq. 11 removes the bias at no cost, and the same normalization argument yields the corrected fingerprint descent $-\nabla_\theta d / \|\nabla_\theta d\|^2$ used for R_{resp}.

B Path Instantiations of Physics-Aware Constraints

PAC is one supervision layer instantiated over four pipelines: free form deformation, neural implicit surfaces, parametric latent vectors, and generative sampling. The two readouts of Section 4.2 never change. What changes is the interface at which each constraint attaches and the form it takes there. This appendix gives each pipeline, its loss and its attachment points, then the rule that organizes them (Section B.5), and closes with the pseudocode of the two differentiable paths.

B.1 Free Form Deformation

Pipeline. Control offsets θ define a Bernstein displacement $\delta(x) = \sum_i B_i(x)\,\theta_i$. The deformability mask multiplies it pointwise, $\tilde{\delta} = m\,\delta$, and the deformed surface enters the frozen surrogate, whose fields integrate into the objective J. The geometry map is linear and pointwise local, so gradients reach every surface point individually.

Loss. With per region budgets $b_r = \kappa\, q_{90}(w)$,

$$\mathcal{L}(\theta) \;=\; J \;+\; \lambda_1 |J_0|\,\mathrm{relu}\big(d/\tau - 1\big)^2 \;+\; \beta \sum_r \Big[\mathrm{relu}\Big(\tfrac{\max_x \|\tilde{\delta}\|}{b_r} - 1\Big)^2 + \mathrm{relu}\Big(\tfrac{\overline{\|\tilde{\delta}\|}}{b_r} - 1\Big)^2\Big] \;+\; \gamma\,\Omega(\theta), \tag{12}$$

where the budget clamps both the peak and the mean displacement of each region and Ω smooths the control net. R_{cover} vanishes identically: the mask has already zeroed the displacement of locked regions.

Attachment. The mask gates the gradient that trains θ,

$$\nabla_{\theta_i} J \;=\; \sum_x m(x)\, B_i(x)\, \frac{\partial J}{\partial p(x)}, \tag{13}$$

so a locked region sends no gradient to any control point. The fingerprint acts as the differentiable penalty of Eq. 5 on the deformed surface, and once more at delivery, where only in band iterates may be returned.

B.2 Neural Implicit Surfaces

Pipeline. A latent code z decodes to a signed distance field, marching cubes (Lorensen & Cline, 1987) extracts the surface, and barycentric samples carry the objective. Their motion with z follows the corrected level set sensitivity $-\nabla f/\|\nabla f\|^2$ of Section 4.3. Fingerprints of marching cubes surfaces live in a different sampling domain than the raw family clouds, so the trust boundary τ is recalibrated on that domain.

Loss. With a ramped fence weight $\lambda(t)$ and a weak code prior,

$$\mathcal{L}(z) \;=\; J \;+\; \lambda(t)\,|J_0|\,\mathrm{relu}\big(d/\tau - 1\big)^2 \;+\; \lambda_r\, R_{\text{cover}}(z) \;+\; \varepsilon\|z\|^2, \tag{14}$$

with R_{cover} the one way Chamfer anchor of Eq. 8 and λ_r calibrated once against the objective's gradient scale. The hinge is silent inside the band, so the fence never pulls the code toward the family mean, and the ramp lets early steps explore and later steps settle.

Attachment. The code is global, so no interface exists at which a local displacement could be zeroed. The deformability field therefore softens from a gate to an anchor. The fingerprint still acts as a penalty because the decoder is differentiable. The fence gradient factors as $\partial R/\partial z = (\partial R/\partial d)\,(\partial d/\partial P)\,(\partial P/\partial z)$. The frozen model reports how each surface point should move for the fingerprint to return to the family, and the level set sensitivity pools those motions into one code direction.

B.3 Parametric Latent Vectors

Pipeline. The design is a low dimensional vector decoded by a procedural parameterization into the hull. The surrogate predicts the objective from the vector directly. The decoder is not differentiable, so no gradient reaches the geometry: the last differentiable interface is the code space itself, and both constraints attach there as gates on each step.

Loss and attachment. Each descent step is decoded, fingerprinted under the family's calibration and accepted only if $d/\tau \le 1$, otherwise it backtracks toward the last accepted iterate by halving until admitted, or stays where it is, so the fence is enforced by acceptance rather than by penalty. The where question is answered by the parameterization's own structure: categorical dimensions are frozen, continuous dimensions are clamped to the family's training range, and the parametrization's feasibility set, forty nine analytic naval constraints evaluated in under a millisecond, is checked at every iterate. Delivery is the surrogate's best iterate among those feasible and in band, and the wave drag oracle enters only there, to measure the delivered hull (Appendix F.3). The same fingerprint gates decoded hulls in the generative use below.

Table 4: One supervision layer, four attachments. The interface at which each constraint attaches, and the form it takes there, follow the path.

	Free form deformation	**Neural implicit surface**	**Parametric latent vector**	**Generative sampling**
Design variable	control offsets θ	latent code z	design vector	sampled candidates
Geometry map	linear, pointwise local	decoder, differentiable, global	procedural, no gradient	decode only, no descent
Fence (how far)	penalty on the deformed surface	penalty through the decoder	gate on each decoded step, $d/\tau \leq 1$	hard gate $d/\tau \leq 1$
Deformability (where)	mask zeroes displacement and gradients	coverage anchor R_{cover}	frozen dimensions, training range, feasibility set	generator support
Enforcement	loss terms	loss terms	accept or reject each step	filter on the stream
Delivery	best feasible iterate	best feasible iterate	best feasible iterate, oracle measured	accept on real

B.4 Generative Sampling

Pipeline. A generator is trained on the family with its native loss, and PAC never touches it. For design vectors this is a tabular diffusion model trained by noise prediction, and for a structural family an occupancy autoencoder trained by reconstruction with a latent prior. At use time the generator proposes candidates, each decoded to a geometry.

Attachment. No descent exists, so neither a penalty nor a gate on steps applies, and the constraint takes its remaining form, a hard gate on the candidate stream. Each decoded candidate is fingerprinted under the family's calibration and kept only if $d/\tau \leq 1$, a zero label filter between decoding and ranking. Survivors are ranked by the surrogate and, when an analytic evaluator exists, accepted on its real feasibility and objective rather than on the surrogate's claim. The gate filters rather than guides: the sampler is never steered by the fence, so the population it proposes is not pulled toward the family mean.

B.5 Why the Modes Differ

Three regularities organize Table 4. First, the attachment follows the last differentiable interface: a penalty where gradients reach the geometry, a gate on each step where they stop at the code, a filter where there is no descent at all. Second, hardness follows locality: a pointwise parametrization admits exact zeroing, a global code admits only anchoring, a candidate stream admits only acceptance or rejection. Third, the pretrained model itself is indifferent to the path: one fingerprint, one deformability statistic, one calibration per family, clamped wherever the pipeline exposes an interface.

B.6 Pseudocode

Algorithms 1 and 2 give the two differentiable paths of Figure 3. Calibration and delivery are shared.

C Implementation Details

This appendix records the four design families, the shared calibration and per path optimization settings, and the analytic oracle behind every verified number.

C.1 Datasets

Figure 8 shows twelve members of each design family.

Algorithm 1 PAC on free form deformation

Input Data: family $\{S_i\}$, initial shape S_0, surrogate f_ϕ, physics geometry model g, both frozen
Input Config: steps T, step size η, weights λ_1, β, γ
Output: trusted design $S^\star$
// Step 1: Calibrate the trust band, label free
$\{z_i\} \leftarrow$ Fingerprint$(\{S_i\}; g)$ // 4096×4 probes.
$d(\cdot) \leftarrow$ NearestFive$(\cdot\,; \{z_i\})$ // Standardized.
$\tau \leftarrow$ Quantile$_{95}(d,$ held out$)$ // How far.
$m, \{b_r\} \leftarrow$ Deformability$(\{S_i\})$ // Where.
Blue marks readouts of the frozen model g.
Initialize $\theta \leftarrow 0$; $\mathcal{H} \leftarrow \emptyset$
for t in $\{1, \ldots, T\}$ **do**
  // Step 2: Deform under the mask
  $\tilde{\delta} \leftarrow m \odot B\theta$ // Locked regions get zero gradient.
  $S_t \leftarrow S_0 + \tilde{\delta}$
  // Step 3: Descend the objective of Eq. 12
  $\mathcal{L} \leftarrow J + \lambda_1 \text{relu}\big(d(g(S_t))/\tau - 1\big)^2 + \beta B + \gamma\Omega$
  $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}$ // g supervises the gradient.
  Append (S_t, J_t, d_t) to $\mathcal{H}$
end for
// Step 4: Deliver only what the band certifies
$S^\star \leftarrow \arg\min\{J_t : d_t \le \tau\}$ // Best in band iterate.
return $S^\star$ // Report claimed and trusted.

Algorithm 2 PAC on neural implicit surfaces

Input Data: family $\{S_i\}$, distance decoder $F(x, z)$, initial code z_0, f_ϕ and g frozen as in Alg. 1
Input Config: steps T, step size η, ramp $\lambda(t)$, λ_r, ε
Output: trusted design $S^\star$
// Step 1: Calibrate on the reconstruction domain
$\{\bar{S}_i\} \leftarrow$ MarchingCubes$(F(\cdot, z_i))$ // Recalibrated domain.
$d(\cdot), \tau \leftarrow$ as in Alg. 1, on $\{\bar{S}_i\}$
Initialize $z \leftarrow z_0$; $\mathcal{H} \leftarrow \emptyset$
for t in $\{1, \ldots, T\}$ **do**
  // Step 2: Decode the current surface
  $S_t \leftarrow$ MarchingCubes$(F(\cdot, z))$
  $P_t \leftarrow$ Barycentric(S_t)
  $\partial P_t/\partial z \leftarrow -\frac{\nabla_x F}{\|\nabla_x F\|^2}\,\partial_z F$ // Corrected sensitivity.
  // Step 3: Descend the objective of Eq. 14
  $\mathcal{L} \leftarrow J + \lambda(t)\text{relu}\big(d(g(P_t))/\tau - 1\big)^2 + \lambda_r R + \varepsilon\|z\|^2$
  $z \leftarrow z - \eta\nabla_z\mathcal{L}$ // Chain rule to z.
  Append (S_t, J_t, d_t) to $\mathcal{H}$
end for
// Step 4: Deliver only what the band certifies
$S^\star \leftarrow \arg\min\{J_t : d_t \le \tau\}$ // Best in band iterate.
return $S^\star$ // Report claimed and trusted.

Blended wing body. BlendedNet (Sung et al., 2025) provides 999 blended wing body geometries simulated with RANS and the Spalart–Allmaras model (Spalart & Allmaras, 1992) at roughly nine flight conditions each, with pointwise pressure and skin friction coefficients. Our processed split carries 1347 training, 153 validation and 1500 test cases. Surfaces are point clouds of about 74k points with normals, per point areas, and three integrated coefficients. The family drives the neural implicit path of Tables 1 and 2 and the free form deformation row of Figure 4.

CRM. The NASA Common Research Model (CRM) (Vassberg et al., 2008) in the control surface study of Catalani et al. (2025): transonic RANS on the deformed flight shape with parametric inboard and outboard aileron, elevator and tailplane settings alongside Mach number and incidence. We use the 44 held out configurations, resampled to 16384 point surfaces with pressure and friction channels. The geometries differ mainly by aeroelastic wing deflection.

Car. DrivAerNet++ (Elrefaie et al., 2024) contains 8000 car designs spanning fastback, notchback and estateback rear configurations, with meshes, parametric models, aerodynamic coefficients and surface fields from high fidelity CFD. Our experiments use 20 unique geometries resampled to 16384 point surfaces with pressure and wall shear channels. The family enters the free form deformation comparisons of Table 1 and Figure 4.

Ship hull. Ship-D (Bagazinski & Ahmed, 2023) describes 30000 hulls by a 45 parameter hull parametrization with wave drag labels. We use its constrained randomized subset of 6000 design vectors as the family. The parametrization's analytic machinery evaluates the Michell integral (Michell, 1898) and the naval feasibility of any candidate, which supplies the only oracle in this paper (Appendix C.3). About 4.5% of generated surface meshes are degenerate and are removed by a width gate before any bank is built. The bank is the set of family shapes whose fingerprints the distance d is measured against, as distinct from the held out shapes on which τ is set (Table 5). The family drives the parametric latent path and the selection experiments of Figures 6 and 7.

C.2 Configuration

Table 5 lists the shared calibration and the per path optimization settings, and every number is recorded with the corresponding run. GeoPT is publicly released, with code at `Physics-Scaling/GeoPT` and the pretrained checkpoint distributed under the MIT licence on Hugging Face. We use its released eight layer model frozen throughout. Both self-calibrated quantities

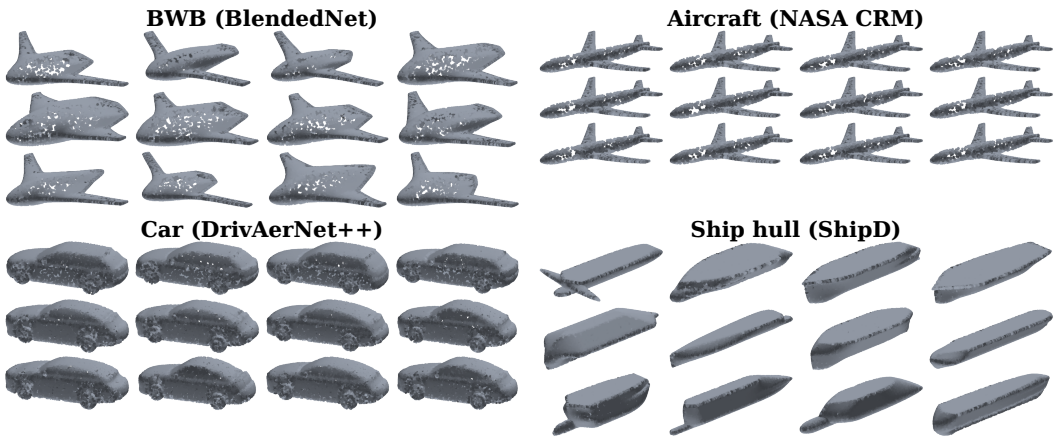


Figure 8: Twelve members of each design family, drawn from the data behind our experiments and rendered with the viewpoints of the main text figures.

of Section 5.2 are read from the family. Against the common rule of a fixed fraction of the geometry, the amplitude budget of Eq. 6 is 1.81 times larger on the car and 4.1 times smaller on CRM, where substituting the fixed rule removes every feasible iterate while the family read budget delivers inside the band. Two entries of Table 5 keep the fingerprint distance d comparable across shapes. The probe region is sized per family, since a box fixed to one dataset falls outside the deformations of another. Fingerprints are compared at matched point density, since a density mismatch alone inflates d by factors that mimic a true excursion. The detection and degradation experiments of Table 3 use 80 held out shapes. In every comparison the unconstrained arm receives the larger of the two displacement budgets. The mask threshold of Section 4.2 is likewise set against the estimation noise floor rather than a quantile of w. Because w is long tailed, a quantile denominator damps the entire surface instead of locking the still regions.

Table 5: Configurations for PAC experiments. Calibration is shared by every path. Optimization follows the path, under Adam (Kingma & Ba, 2015).

(a) Calibration configuration.

Terms	Configure
Response fingerprint	
Probe points	4096 per set
Probe sets	4
Probe box	family adaptive
Distance d	mean top-5 neighbor
Standardization	per dim z score
Trust boundary τ: p_{95} of held out	
BWB, neural implicit	bank 113, held 40
BWB, FFD	bank 32, held 24
CRM	bank 28, held 15
Car	bank 12, held 7
Ship hull	bank 400, held 40
Deformability field	
Estimator	point to tangent plane
Mask free level k	$1.5\, w_0$
Region budgets b_r	$1.5\, q_{90}$ of variability
Labels used	**none**

(b) Optimization configuration.

Terms	Configure
Free form deformation path	
Optimizer	Adam, fixed step
Steps	120
Fence weight λ_1	2.0, in units of $\lvert J_0 \rvert$
Step size, per family	budget / $\{8, 320, 40\}$
Smoothness $\gamma\, \Omega$	not used
Neural implicit path	
Optimizer	Adam, annealed
Steps	180, logged every 5
Fence weight	ramped to 2.0
Parametric and generative paths	
Candidates sampled	1400
Kept by the fence	4
Oracle quadrature	Michell, $N = 16$
Delivery, all paths	**best in band iterate**

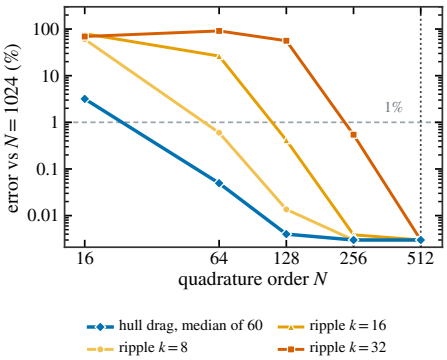
C.3 The Analytic Oracle

Michell's thin ship theory (Michell, 1898) expresses the wave resistance of a slender hull as a double integral of the hull's longitudinal slope over the wetted surface and over wave propagation angles. The Ship-D parametrization (Bagazinski & Ahmed, 2023) evaluates it on any design vector, together with its naval feasibility gates. We report it as the coefficient $C_w = R_w/(\frac{1}{2}\rho L^2 U^2)$, the wave resistance non dimensionalized by the dynamic pressure and the square of the length overall, at the speed the family's Froude number sets. Two properties make it the reference judge of this paper. It is analytic, so any candidate can be scored without a solver run. It is independent of every learned component and is never queried by the optimizer, so no optimization pressure can reach it. It enters at delivery only, to certify feasibility and measure true wave drag.

Convergence. Both discretizations of the integral are refined until the reading stabilizes. Table 6 sweeps the wave angle quadrature against an $N = 1024$ reference. The two uses of the oracle converge at very different orders. Hull selection is insensitive almost immediately. At $N = 16$ the ranking of 60 hulls already correlates at 0.997 with the sixteen kept hulls identical, and from $N = 64$ the median reading moves by 0.05% or less. The fine ripple probes of Appendix E, Gaussian windowed surface wavelets of identical displacement (Figure 11), are the demanding client. At $N = 16$ their drag response is off by 60 to 79%, and the $k = 32$ probe aliases non monotonically through $N = 128$. Every probe agrees with the reference to better than 0.01% only from $N = 512$, the setting the probes use. Hull scoring runs at $N = 16$, where the same sweep shows the kept set and the ranking already converged. The spatial grid is refined likewise, until the integral moves by less than 1%.

Table 6: Convergence of the Michell quadrature against $N = 1024$. Left: ranking correlation, kept set overlap and median drift of 60 hulls, and the response error of three ripple probes. Right: the error of each reading against the $N = 1024$ reference. The dotted line marks $N = 512$, the order the ripple probes use. Hull scoring runs at $N = 16$, where the ranking is already converged.

	Hull selection			Probe error (%)		
N	ρ	top-16	med. $\lvert\Delta\rvert$	k=8	k=16	k=32
16	0.997	16/16	3.18%	60.4	79.4	69.2
64	1.000	16/16	0.05%	0.6	26.4	91.2
128	1.000	16/16	<0.01%	<0.1	0.4	56.1
256	1.000	16/16	<0.01%	<0.1	<0.1	0.5
512	1.000	16/16	<0.01%	<0.01	<0.01	<0.01



D Full Results and Repeated Runs

The single run claims of Section 5 are repeated here over five query geometries on the blended wing body and the car, or over three starting points and two seeds on the neural implicit path. The full versions of the two main text tables are given alongside.

Full result tables. Tables 7 and 8 give the complete versions of the two main text tables. The first adds the parametrization of each family and the footnotes on the ship hull row. The second adds the claimed optimum of each factorial arm alongside the trusted one.

Factorial replication. Over three further starting points and two seeds (Figure 9), the unconstrained arm ends outside the trust band in six runs of six, at 3.2 to 10.3 times the boundary. The deformability field only arm does likewise, at 1.0 to 9.5. Neither terminal claim is deliverable. The fence only and combined arms deliver inside the band in six of six. The median share of the search spent inside the band rises from 11% unconstrained to 14% and 44% under each constraint alone, and to 67% under both. Table 9 lists the delivered outcome of every run. Each constrained arm delivers inside

Table 7: Inverse optimization with and without PAC on four CAE families, the full version of Table 1. Claimed is read from the surrogate and trusted is the best iterate inside the band. d/τ is the distance to the family response manifold in units of the trust boundary. Only the ship hull carries an analytic oracle.

		Unconstrained		**PAC (Ours)**	
Family	Parametrization	claimed	d/τ	trusted	d/τ
Blended wing body	neural implicit	$+55.7\%$	11.74	$+5.1\%$	0.86
NASA-CRM	free form deformation	$+1290.7\%$	639	$+2.2\%$	0.85
Car	free form deformation	$+526.8\%$	12.69	$+22.7\%$	0.70
Ship hull	parametric latent	infeasible†	1.64	$+64.5\%^{\ddagger}$	0.47

† The surrogate claims the wave drag vanishes ($\log_{10} C_w = -8.89$), but the analytic oracle rules the hull infeasible. ‡ *Verified*: measured by the analytic oracle rather than read from the surrogate, under the protocol of Appendix F.3.

Table 8: Factorial dissociation of the two constraints on the neural implicit path, the full version of Table 2. Each arm removes one constraint. Claimed is the surrogate's best report along the trajectory, and trusted is the best iterate inside the band, undefined without the fence. Terminal d/τ is where the search ends, and in band counts the logged iterates inside the band, of 36. Bold is best per scored column.

		Claimed optimum		Trusted optimum		Search	
R_{resp}	R_{cover}	improvement	d/τ	improvement ↑	d/τ	terminal d/τ ↓	in band ↑
—	—	$+55.7\%$	11.74	/	/	11.74	2
✓	—	$+19.8\%$	5.44	$+8.8\%$	0.82	1.10	6
—	✓	$+5.1\%$	0.86	$+5.1\%$	0.86	1.62	16
✓	✓	$+5.1\%$	0.86	$+5.1\%$	0.86	**0.97**	**24**

the band from every start and seed, while the unconstrained claim never lands closer than 1.95 times the boundary. The two constraints divide the axes of trust, not the magnitude of the report. The pair delivers the field arm's iterate in six of the seven runs and ends the search at 0.85 to 2.74 times the boundary, a median of 1.11 against 1.62 and 1.72 for each constraint alone, inside the band in three runs of seven where neither alone ever does. The arms separate on where the search ends, which Table 8 and Figure 9 measure, and on which regions it disturbs, which Table 10 measures. All seven runs share one calibration card, one λ and one script revision, recorded with each result.

Table 9: Delivered outcome of every factorial run: the run of Table 2 (A/0) and three further starting points under two seeds. Claimed and trusted are as in Table 8. The d/τ of every trusted delivery is at most 1.00.

	Unconstrained		R_{resp} only		R_{cover} only		**Both (PAC)**	
Start/seed	claimed	d/τ	trusted	d/τ	trusted	d/τ	trusted	d/τ
A/0	$+55.7\%$	11.74	$+8.8\%$	0.82	$+5.1\%$	0.86	$+5.1\%$	0.86
B/0	$+46.1\%$	7.71	$+14.6\%$	0.76	$+9.3\%$	0.73	$+9.3\%$	0.73
B/1	$+60.4\%$	5.91	$+19.2\%$	0.83	$+7.6\%$	0.71	$+7.6\%$	0.71
C/0	$+183.5\%$	8.51	$+12.4\%$	0.84	$+10.3\%$	0.99	$+10.3\%$	0.99
C/1	$+160.1\%$	10.20	$+26.4\%$	1.00	$+31.2\%$	0.97	$+30.6\%$	0.86
D/0	$+40.7\%$	1.95	$+28.2\%$	0.76	$+4.1\%$	0.68	$+4.1\%$	0.68
D/1	$+54.2\%$	3.65	$+0.4\%$	0.68	$+0.4\%$	0.68	$+0.4\%$	0.68

Local edits and the pointwise constraint. The dissociation of Section 5.1 also appears under free form deformation, measured directly by imposing deformations of one fixed amplitude on the car family (Table 10). The fence reads one global distance, so a sufficiently local edit leaves it unmoved. Concentrated on 0.3% to 5% of the surface, the deformation leaves the reading at 0.25 to 0.27, indistinguishable from the initial shape. The same amplitude spread over the whole surface moves it

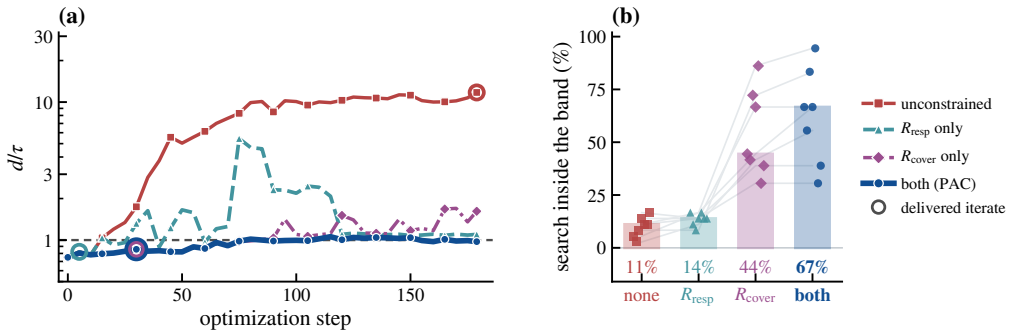


Figure 9: Factorial arms along the optimization trajectory. (a) Distance to the family manifold in units of the trust radius τ, for the starting point of Table 2. The dashed line is the trust boundary and rings mark the delivered iterate, concentric where two arms deliver the same one. (b) Share of the 36 logged iterates that lies inside the band, one thin line per run over three starting points and two seeds. Bars are medians.

to 1.26, out of the band. The deformability field reads the same edits the other way around. It assigns the local sites lower weight than the globally spread one, 0.82 to 0.87 against 0.96, so the mask damps exactly what the fence cannot see. The pair covers both scales, which is the orthogonality Table 2 measures.

Table 10: Deformation probes of one fixed amplitude on the car family. The fence reading separates global from local, and the deformability weight at the deformed sites separates local from global.

Deformation	surface moved	fence d/τ	field weight at site
none (initial shape)	0%	0.25	/
local	0.3%	0.25	0.87
local	1%	0.25	0.84
local	5%	0.27	0.82
global	100%	1.26	0.96

Locked regions under free form deformation. The free form deformation path measures how well an arm protects the regions the family holds fixed, with exact point correspondence and a locked set of 1.24% of the surface. On the car, with the same budget and optimizer in every cell, the ratio of locked to free displacement is 0.728 unconstrained, 0.934 with the fence alone, 0.310 with the field alone and 0.331 with both, a lower ratio meaning the locked region moved less. Each reading lies outside a placebo band built from random equal sized subsets of the free points. The fence reads one global distance and leaves the locked region to the search, while the field pins it, the division of labor Table 10 measures probe by probe.

E Additional Analysis

What pretraining adds to the backend. Figure 10 reads Table 3 along its two axes at once. The generative encoder spends more than its whole trust band on a harmless 0.3% resampling noise, on every held out shape. The physics lifted readout spends a fifth of its band and detects a 2% deformation at the highest AUC. The voxel descriptor sits between them, and Appendix F.4 shows that where it sits is a property of the configuration it is read under rather than of the descriptor. Panel (b) gives, per band, the net effect of pretraining reported in Section 5.2. The oracle column of Table 3 composes differently: the selection is carried by the distance field representation and the calibration protocol, and the contribution of pretraining lies on the detection axis.

Position and scale sensitivity. Figure 5 resolves the induced metric along one axis at a time, and Figure 11 resolves both at once. Localized wavelet probes sweep position along the hull and deformation wavelength. Every probe carries the same geometric displacement, so the maps report only where and at what scale a constraint responds. The three hot regions do not overlap: the true

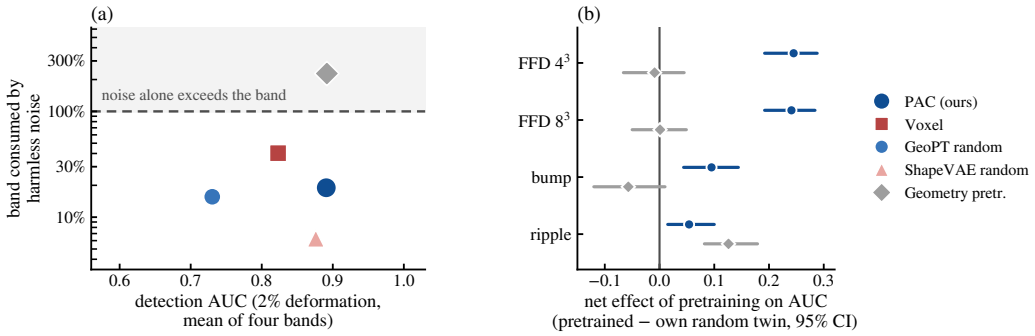


Figure 10: Composition of the trust backend, visualizing Table 3. (a) Detection of a 2% deformation against the share of the band consumed by harmless noise. Above the dashed line the noise alone exceeds the band. (b) Net effect of pretraining per deformation band, with 95% intervals.

wave drag concentrates at long wavelengths near the Kelvin scale, geometry only pretraining at short wavelengths, and physics lifting is nearly uniform. Measured as peak over mean, the concentration is 2.03 for the true wave drag, 1.50 for physics lifting and 2.26 for geometry only pretraining, whose random twin sits at 1.47. Figure 12 reads the same maps as a charge ratio against the true wave drag sensitivity. Geometry only pretraining overcharges the finest wavelengths by up to 8.6× and undercharges the Kelvin band where the drag lives. Physics lifting stays closest to parity, its median charge within 4% of the true sensitivity against 15%.

The horizontal axis delimits what any global fingerprint can read: a fingerprint answers how far and at what scale, not where along the hull a fixed deformation matters. PAC assigns the where question to the pointwise deformability field for exactly this reason (Section 4.2). The two constraints thus split the two axes of this map between them. Cells rejected by the displacement guard, 1.7% of the total, are lost identically by every backend.

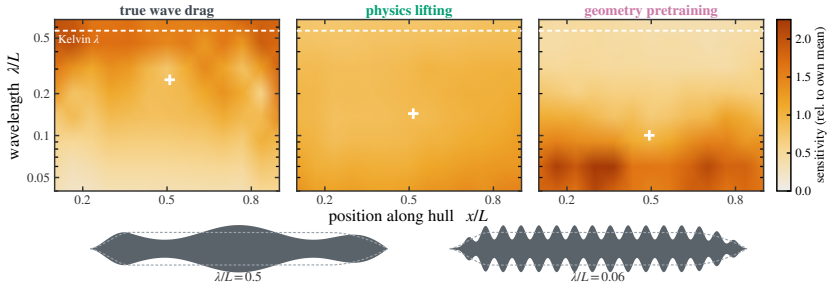


Figure 11: Position and scale resolved sensitivity of each judge. Axes are position along the hull and deformation wavelength, and colour is the response normalized to each map's own mean. Probes are Gaussian windowed wavelets of identical displacement. The dashed line is the Kelvin wavelength $2\pi Fn^2$ and the cross is each centroid.

Trust regions under free form deformation. Figure 13 repeats the construction of Figure 6 on the two free form deformation families. The vertical axis follows the displacement the unconstrained arm of Figure 4 delivered. On this plane the fence readings at the delivered designs reproduce those of Figure 4 to within 1.2%. The family projects onto a thin strip around the chord. Every fence rejects the delivered exploit, at 2.4 to 8.4 times its boundary on the blended wing body and 17 to 2.5×10^4 on CRM. The fences differ in the shape of what they keep. On the blended wing body the voxel descriptor admits a slab that extends halfway toward the exploit regardless of direction. The PAC band hugs the family strip, keeping 88% of the members in a region 3.7 times smaller, and rejects the exploit the deepest. On CRM, whose trust band is the narrowest of the four families, only the physics lifted band still contains the family. It holds 17 of 20 members, against 1 for the voxel descriptor and 0 for the pretrained encoder. The PAC delivery projects inside the band on both families.

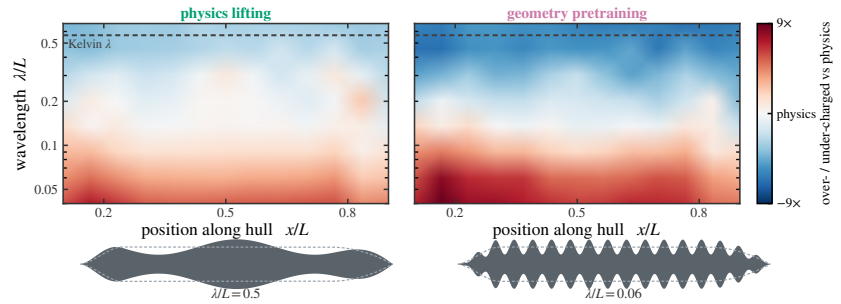


Figure 12: Charge ratio of each pretrained backend against the true wave drag sensitivity, on the maps of Figure 11. Red overcharges and blue undercharges. Insets show the probe deformation at two wavelengths.

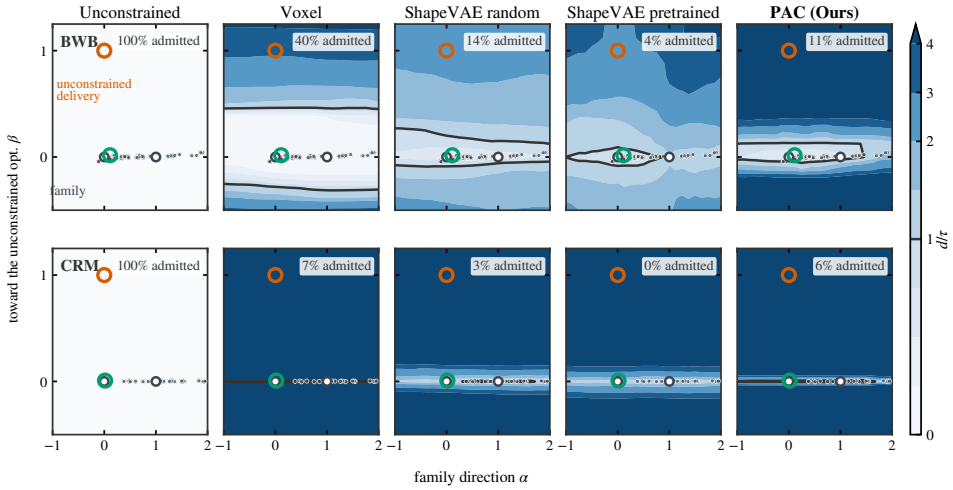


Figure 13: Trust regions on one plane of each free form deformation family, constructed as in Figure 6 with the vertical direction following the unconstrained delivery of Figure 4 (ring). The green ring is the PAC delivery. Colour is d/τ under each backend's own calibration. Dots are family members, and annotations give the share of the plane admitted.

F External Trust Controls

This appendix answers four questions about the trust band, each with one measurement: whether the surrogate is more accurate inside it, and whether a distance in the design parameter space or the surrogate's own uncertainty could replace it. It then separates the delivery protocols on the ship hull and measures how much of a geometry descriptor's cost is its configuration.

F.1 Does the Band Track Surrogate Error

The band marks where surrogate error rises and does not rank error inside the family, because that error is set by the operating condition rather than the geometry. On the blended wing body, where every case carries a CFD label, a surrogate of the production architecture trained on the 24 bank geometries alone errs 12.6% outside the band and 8.2% inside, read on the 129 geometries it never saw. Inside the family a random effects decomposition of the per case error attributes eight per cent of its variance to the geometry, an intraclass correlation of 0.08 that bounds any geometry only predictor near $\rho = 0.29$. The band is therefore a support indicator, and the trusted level of evidence is reported apart from the verified one.

F.2 A Fence in the Design Parameter Space

A distance in the design vector is the cheapest fence a practitioner has (Eriksson et al., 2019; Toal et al., 2010). On the ship hull, the one family with such a space, it keeps hulls 0.8 to 3.3 times below the pool median at $K = 16$, where the fingerprint keeps 10.9 (Table 11), under the protocol of Section 4.2 with only the vector changed. Such a fence rejects corrupted vectors, at 95 to 100% under one standard deviation of noise per dimension, but it has nothing to reject among sampled candidates. Between 99 and 100% of them lie inside its τ, since a sampler trained on the family emits vectors that are in distribution in that space. The wave drag is carried by the hull's proportions, which the fingerprint recovers at $\rho = +0.73$ and the parameter distance does not, at $+0.06$. Such a fence also exists only where the family shares one fixed parametrization, whereas the fingerprint takes a surface and applies unchanged to all four families. The ship hull path therefore runs the fingerprint fence in the loop (Appendix B.3), and Appendix F.3 shows that a parameter band at the family radius admits hulls the fingerprint places 4.8 times outside its band.

Table 11: A fence read on the design vector against the response fingerprint, on the ship hull. Oracle gain is the pool median true wave drag divided by the median over the K hulls each fence keeps.

	Oracle gain ↑ at K kept				
Fence measured on	K=4	K=8	K=16	K=32	$\rho(d, C_w)$ ↑
Design parameter space, the 45 dimensional hull vector					
nearest neighbor, standardized	2.4×	0.9×	0.8×	1.1×	+0.061
whitened, to the family mean	4.1×	2.4×	3.3×	1.9×	+0.115
whitened, nearest neighbor	2.8×	2.3×	2.8×	1.4×	+0.084
orthogonal decomposition, 95%	2.8×	4.5×	3.1×	1.9×	+0.041
Response fingerprint, GeoPT (Ours)	**17.4**×	**13.1**×	**10.9**×	**6.5**×	**+0.726**

F.3 Delivery Protocols on the Ship Hull

Once the naval feasibility set is read at every iterate, selecting the delivered hull by the surrogate or by the oracle returns the same hull from every start, and so does every fence. The optimization of Table 1 is repeated from the start of Figure 1 and twelve further feasible family hulls, with the same surrogate, descent, frozen dimensions and training range. Every iterate is recorded with its fingerprint distance, its distance to the nearest family vector, its feasibility and its true wave drag. Four fences bound the steps: none, a parameter band at 0.4 and at 1.0 of the family's 90th percentile nearest neighbor radius r_{90}, and the fingerprint fence as a gate on each decoded step. Three delivery rules differ only in what the oracle may do (Table 12). The surrogate's descent leaves the feasible set within a median of four steps. A band bounds drift and not feasibility, so it delivers an infeasible hull whenever the oracle only measures, which is why the parametric path reads the feasibility set at every iterate (Table 4). With feasibility read, the surrogate's ranking inside the band carries the verified improvement, 24% at the median and 64.5% from the start of Figure 1. The fences separate on where the search ends. Unconstrained, it ends 1.4 family radii from the nearest family vector, and under the parameter band at hulls the fingerprint places up to 4.8 times outside its band. Under the fingerprint fence it ends at $d/\tau \leq 1.00$ on every start, after a median of 159 of 200 steps are rejected. Two further starts lie at or beyond the boundary, at d/τ of 1.01 and 1.30, and the fence admits the first step of one and returns the other unchanged.

F.4 How Much of a Descriptor's Cost Is Its Configuration

Most of its cost is the configuration, and the frame decides more than the resolution. Across thirty configurations of the voxel descriptor, one framed on the reference bank inherits whatever the bank holds. On the ship hull the 26 degenerate meshes the width gate removes cost it five sixths of its selection power, from 11.9 to 2.4 times below the pool median. Framed per shape it is unchanged, as is the fingerprint. We normalize per shape for this reason, and all reported numbers are gated (Appendix C.1). Smoothing the occupancy grid brings its noise cost inside the band. What no configuration reaches is the detection axis: paired over the same held out shapes and seeds, the

Table 12: Delivery protocols on the ship hull, over the eleven starts inside the trust band. Terminal is the last iterate, with its distance to the nearest family vector in units of r_{90}. First infeasible is the median step at which the descent leaves the naval feasible set. The three delivery rules read the same trajectories with the oracle measuring only, gating feasibility, or selecting the hull, and report the deliveries the oracle certifies feasible and the median verified improvement in wave drag.

	Terminal			Measure	Gate		Select
Fence	feasible	distance	First infeasible	feasible	feasible	gain	gain
None, training range only	$0/11$	1.42	4	$0/11$	$11/11$	$+24.2\%$	$+24.2\%$
Parameter band, $0.4\,r_{90}$	$0/11$	0.40	4	$0/11$	$11/11$	$+24.2\%$	$+24.2\%$
Parameter band, $1.0\,r_{90}$	$0/11$	1.00	4	$0/11$	$11/11$	$+24.2\%$	$+24.2\%$
Fingerprint fence (Ours)	$0/11$	0.65	4	$0/11$	$11/11$	$+24.2\%$	$+24.2\%$

physics lifted readout beats every voxel configuration by $+0.067$ to $+0.130$ mean AUC, with all eleven intervals excluding zero. The share of shapes on which a harmless 0.3% resampling crosses the whole band, a statistic independent of where a heavy tailed median falls, separates the backends the same way. It is every held out shape for the generative encoder, 39% for the voxel descriptor and none for GeoPT.

F.5 The Surrogate as Its Own Fence

The surrogate's own uncertainty cannot serve as the fence, because a judge trained on the data that trained the exploited model inherits its blind spot. Five members of the production architecture trained on the bank geometries alone, with a sixth carrying dropout for a Monte Carlo estimate, predict the drag coefficient at $R^2 = 0.98$ on held out geometries and rank their own error at $\rho = +0.55$. Read with normals re-estimated as the geometry moves, as the inverse loop does, the best ensemble statistic detects a 2% deformation at a mean AUC of 0.65 against 0.89 for the frozen pretrained readout, with paired gaps of $+0.087$ to $+0.344$ that exclude zero. Dropout sits at chance on every band. The members stop disagreeing where disagreement is needed: under one perturbation their predicted changes correlate at 0.72 at one per cent of amplitude and at 0.93 at forty, while their collective drift grows from 1.6 to 3.6 times the spread between them. This is why PAC reads its constraint from a model that was never fitted to the family.

G Limitations and Future Work

PAC certifies that a delivered design stays where the surrogate carries evidence. It certifies neither optimality within that region nor a verified improvement where no oracle exists. The band certifies support rather than accuracy: on the blended wing body the surrogate errs about half again more outside the band than inside it, and the error that remains inside is set by the operating condition rather than the geometry (Appendix F.1), which is where a band would have to look next. Within the region the surrogate's own ranking selects the iterate, since the supervisor is deliberately denied the descent direction (Section 4.1). Verification is withheld because the originating solver of a published dataset cannot be re-run (Section 1). Delivery records every iterate with its distance and claim, so a later re-simulation can upgrade trusted results to verified ones without re-running a search.

PAC does not replace a solver in the loop. A design that matters is still confirmed by simulation before it is built. What the two constraints change is what reaches that step: every delivered design arrives with its distance and its claim recorded, so the verification budget is spent on candidates worth simulating rather than on rejecting exploits.

The fingerprint reads how far a deformation goes and at what scale rather than where (Appendix E). A pretraining objective with localized readouts could let one readout answer both questions. Two further directions follow from the design: a trust aware search that schedules its steps by the distance to the boundary, and, to us the most interesting, a trust boundary extended to the joint space of geometry and operating condition, where a surrogate's competence is decided.