CAELUM: SELF-EVOLVING COMPILATION OF FORWARD–INVERSE CAE OPTIMIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Inverse design is the promise of learned physics surrogates: differentiate through the surrogate, and simulation data becomes better geometry. Realizing it on a new dataset still demands a hand-built pipeline and dozens of expert decisions. The common remedy is to pretrain one large model across regimes. Yet CAE data is *proprietary and rarely poolable*, so specializing to each dataset is the more practical path. We present **Caelum**, a self-evolving compiler for this setting: from a raw CAE dataset it emits one deployable forward–inverse optimizer that predicts physical fields and reshapes geometry through them, specialized to its dataset and never leaving its owner. Our key insight is that build expertise is itself learnable. A simulation label costs a solver run, but an alternative build decision costs only retraining on data already in hand, so counterfactual build sweeps turn fifteen datasets into hundreds of process-rewarded preference pairs. Confirmed lessons harden into rules and precedents, and rule-residual preference optimization (RRPO) then tunes the language model only on decisions no rule can answer. Caelum ships every optimizer end-to-end across fifteen datasets from four physics domains. On frozen decision boards it answers every regime probe that frontier models miss and cuts the decision loss from 22.4 R^2 points to 0.1. Design-loop construction becomes a self-evolving compilation process.

1 INTRODUCTION

Neural surrogates have made physics prediction fast. Trained on solver outputs, they amortize hours of numerical simulation into a single forward pass (Li et al., 2021; Kovachki et al., 2023; Wu et al., 2024), now at industrial scale (Luo et al., 2025; Zhou et al., 2026; Adams et al., 2025). The payoff industry wants sits one step further, in inverse design: wrap an objective around the surrogate and differentiate, and every gradient step reshapes the geometry itself (Remelli et al., 2020; Allen et al., 2022). Yet this capability has not translated into routinely deployed design loops.

Building such a loop remains expert-intensive, because a deployable loop is far more than a trained network. Data must be ingested and split without leakage (Kapoor & Narayanan, 2023). A model family must be picked from a fast-growing zoo (NVIDIA Corporation, 2025) and sized to the physics regime. Objectives must be wired to gradients that actually exist, and constraints must keep the optimization physical. Every one of these is an expert decision, and every new dataset reopens all of them (Hutter et al., 2019). Wrong choices fail silently, because the pipeline still trains and converges while accuracy degrades. On a single dataset in our experiments, the gap between defensible configurations reaches 94 R^2 points. The bottleneck is not the network but the decisions around it.

One remedy is to make these decisions once, by pretraining a physics foundation model that amortizes one network across regimes (Wu et al., 2026). Yet industrial CAE data is held inside company walls and seldom pooled, and even a shared backbone must be specialized before it is deployed. The practical route runs the other way: each dataset a user brings is compiled into a design loop of its own and deployed inside its owner's walls, meeting data scarcity with cheap bespoke construction in place of one wide network. The decisions then recur with every dataset.

Delegating these decisions to a frontier language model does not resolve them. LLM agents automate a growing share of mainstream machine-learning engineering (Chan et al., 2025; Jiang et al., 2025), but CAE builds are different, because the deciding evidence appears in no pretraining corpus. Whether geometry-aware attention (Adams et al., 2025) helps depends on how fields vary. Whether

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.

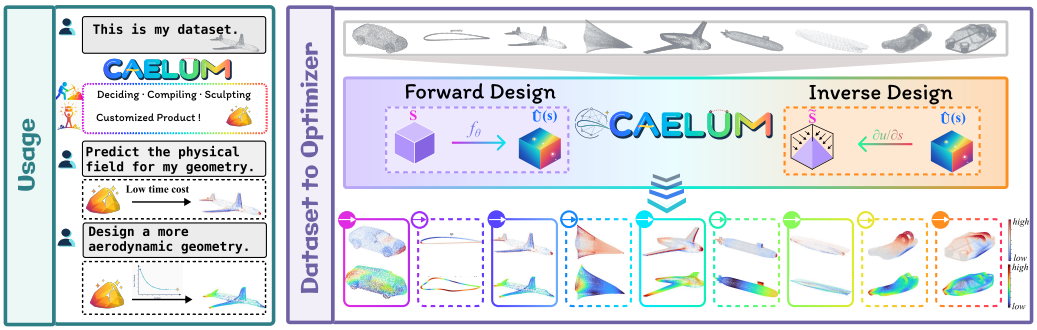


Figure 1: Caelum compiles a raw CAE dataset into a deployable forward–inverse optimizer. Left (Usage): an engineer submits a dataset and receives a compiled product that predicts the field of a geometry and redesigns it. Right (Dataset to optimizer): a surrogate f_θ predicts the field $\hat{u}(s)$ of a shape s, and differentiating it drives the geometry toward an improved design $\tilde{s}$. Bottom: outputs of the products compiled for every dataset.

a wider model wins depends on how many independent geometries the dataset holds. In our benchmark, four of nine zero-shot frontier LLMs mis-select the model family on one structural dataset and forfeit 0.49 R^2 on the resulting product (Appendix C). The knowledge does not exist until the experiment is run.

Experience must therefore be earned, and then retained. The natural remedy is to log build outcomes as preferences and fine-tune the deciding policy (Hu et al., 2022; Rafailov et al., 2023). At CAE scale this remedy breaks down: every preference pair is paid for by a full training run, so the corpus cannot be grown the way an alignment corpus can. On a corpus that small, tuning becomes a competition for capacity (McCloskey & Cohen, 1989; French, 1999). In our own policy generations, each repair evicted another lesson (Fig. 6a in Appendix C). This mirrors recent evidence that forgetting grows with how far tuning data departs from the model's own distribution (Chen et al., 2026a; Zhang et al., 2026a). Consolidating every lesson into the weights is self-defeating.

We treat build expertise itself as the thing to learn. Two asymmetries make that tractable. The first is economic. Once a dataset exists, every alternative build decision costs one more training run on it, whereas every new label costs a solver run. Counterfactual sweeps over build choices multiply one dataset into dozens of causally attributed preference pairs, each scored by a per-decision process reward (Lightman et al., 2024). The second is representational: lessons of every kind compete for the same weights, yet their storage requirements differ (McClelland et al., 1995). A confirmed conditional is exact, and a rule stores it losslessly. A single instance is evidence, and a retrieval library keeps it verbatim (Aamodt & Plaza, 1994). Weights are irreplaceable only for what neither medium offers: smooth generalization to unseen regimes (Silver et al., 2018). The key insight is therefore to route each kind of lesson to its own medium. The tuning corpus shrinks to the residual, stays approximately on-policy by construction (Chen et al., 2026a), and the capacity competition disappears outright.

Caelum operationalizes both asymmetries as a self-evolving compiler. Given a raw dataset, it profiles the physics regime, builds the product, and sweeps its own decisions along the way: every arm it trains becomes evidence, every outcome is scored per decision, and ties are archived instead of discarded. The evidence then settles into a three-tier arbiter. Statutes answer the decisions that confirmed conditionals cover. A precedent library answers by retrieval where instances suffice. The LLM is consulted only where both are silent, open-book, and rule-residual preference optimization (RRPO) tunes it on that residual. Concretely, an engineer submits a directory of solver results. Hours later, a deployed tool answers the two questions a designer actually asks: *what will this shape do*, and *how should it change*. In this study that pattern repeats fifteen times over: dataset in, design loop out. Behind that throughput, the decision loss on the frozen board collapses from 22.4 R^2 points under the pre-lesson defaults to 0.1 in replay (Fig. 6b). Our contributions are threefold:

- To our knowledge, Caelum is the first system in the industrial CAE domain to compile raw datasets into deployed forward–inverse optimizers, recasting design-loop construction as a compilation problem and carrying it out end-to-end across fifteen datasets and four physics domains.
- Counterfactual build sweeps with per-decision process rewards manufacture causally attributed preference pairs from scarce CAE data, with ties and negative results as first-class evidence.
- Case-law consolidation routes each lesson to statutes, precedents, or the weights by the coverage of its evidence. Lessons the statutes hold are out of reach of tuning, and RRPO tunes the weights on the residual alone, reproducing every deployed decision of the monolithic corpus with 19% of its pairs.

2 Related Work

2.1 Differentiable Simulation

Neural simulators. Neural operators formalize simulation as learning maps between function spaces (Kovachki et al., 2023; Li et al., 2021; Pfaff et al., 2021; Liu et al., 2026a). Transolver treats mesh points as tokens and learns latent physical states (Wu et al., 2024). It has since been scaled to million-cell industrial geometries (Luo et al., 2025; Zhou et al., 2026; Adams et al., 2025) and pre-trained on off-the-shelf geometry (Wu et al., 2026). This line is complementary to ours. Caelum consumes these solvers as interchangeable parts through the open PhysicsNeMo zoo (NVIDIA Corporation, 2025), so each solver advance improves the compiled product with no change to Caelum.

Inverse design. Differentiability turns a trained surrogate into a design tool. MeshSDF optimizes shapes through an implicit surface field (Remelli et al., 2020), and learned graph simulators support gradient-based design (Allen et al., 2022). Each is a single loop assembled by experts for one problem. An LLM agent has since automated the assembly for photonic metamaterials, one loop per query (Lu et al., 2025). Our work constructs such loops across CAE datasets and physics regimes and retains the build evidence between them.

2.2 CAE Agents

Automated pipelines. Classical AutoML tunes models within a fixed pipeline (Hutter et al., 2019). LLM agents extend the scope to full engineering workflows: they search the space of solution code (Jiang et al., 2025) and are benchmarked on Kaggle-style competitions (Chan et al., 2025). Both operate where the deciding evidence is public, and both restart with every task. CAE build decisions are instead settled by evidence produced experimentally. Caelum accumulates that evidence rather than rediscovering it.

Agents for simulation. A recent line automates solver-side CFD workflows (Chen et al., 2024; Pandey et al., 2025; Fan et al., 2025; Yue et al., 2025). These agents run individual cases and carry nothing over between tasks. Their simulations are the raw material Caelum compiles.

Self-evolving agents. Agents that improve from their own experience keep lessons outside the weights, as case banks, guidelines, reasoning memories, evolving playbooks, and residual policies (McClelland et al., 1995; Aamodt & Plaza, 1994; Silver et al., 2018; Zhao et al., 2024; Fu et al., 2024; Zhou et al., 2025; Hatalis et al., 2025; Tao et al., 2026; Ouyang et al., 2026; Zhang et al., 2026b; Gao et al., 2026), or write them into the weights by preference tuning (Rafailov et al., 2023; Hu et al., 2022), where small corpora interfere and earlier lessons are forgotten (McCloskey & Cohen, 1989; French, 1999; Lin et al., 2026; Chen et al., 2026a; Zhang et al., 2026a). Over repeated rounds, internalized experience has been found to collapse rather than compound (Chen et al., 2026b). Caelum differs in where its experience comes from and where it goes. Its lessons are manufactured by controlled counterfactual sweeps and scored by solver outcomes, not reflected by the model from its own trajectories. Statutes and precedents take the parts of principle-level and instance-level experience, a statistical gate promotes them, and only the residual reaches the weights. A neuro-symbolic world model that tunes its neural component only on trajectories its symbolic rules fail to explain (Zhao et al., 2026) is the closest mechanism to that partition. RRPO makes it exact, tunes the weights on the residual alone, and recovers both extremes as degenerate cases.

3 PROBLEM SETUP

We consider a CAE dataset $\mathcal{D} = \{(g_i, s_i, u_i)\}_{i=1}^{n}$ pairing geometries g_i and operating conditions s_i with solver-computed fields u_i. A build trains a surrogate $F_\theta(g, s) \mapsto u$ on $\mathcal{D}$ and couples it to a differentiable objective. The result is a forward–inverse optimizer, which returns the field of a geometry and, by differentiating through it, an improved geometry. For example, in aerodynamics g is a wing surface mesh, s is the flight condition, u is the surface pressure, and the compiled optimizer returns a lower-drag wing.

A build is fixed by a sequence of K discrete decisions $a = (a_1, \ldots, a_K)$, where each $a_k \in \mathcal{A}_k$ ranges over one axis such as the model family or the capacity. Executing a on $\mathcal{D}$ yields a product scored by $Q(\mathcal{D}, a)$: held-out accuracy for the forward map, and the attained design objective for the inverse loop. For a policy π that emits $\hat{a}$, let $a_k^\star = \arg\max_{a_k \in \mathcal{A}_k} Q(\mathcal{D}, (\hat{a}_{-k}, a_k))$ be the best choice on axis k with the remaining decisions held at $\hat{a}_{-k}$. The policy incurs the decision regret

$$R(\pi; \mathcal{D}) = \sum_{k=1}^{K} \big[\, Q\big(\mathcal{D}, (\hat{a}_{-k}, a_k^\star)\big) - Q\big(\mathcal{D}, \hat{a}\big) \,\big]. \tag{1}$$

In current practice a human expert supplies a for every new dataset. This paper seeks a policy π that emits a from a descriptor $\phi(\mathcal{D}) \in \mathcal{X}$ observable at ingestion. The descriptor records, for example, the physics kind and the numbers of distinct geometries and operating conditions. The policy should minimize $\mathbb{E}_{\mathcal{D}}[R(\pi; \mathcal{D})]$ without any solver call beyond those already in $\mathcal{D}$.

4 METHOD

Learning the policy π of Eq. 1 requires two things. The first is supervision that reveals which decision wins on which dataset, a signal no existing corpus contains. The second is retention that survives on a corpus far smaller than any alignment set. Caelum obtains the first by counterfactual build sweeps. It obtains the second by consolidating the sweeps as case law and tuning the language model on the residual alone (Fig. 2).

4.1 MANUFACTURING SUPERVISION

Credit assignment. Write $\Delta_k(\mathcal{D}; a_k, a_k') = Q(\mathcal{D}, (a_{-k}, a_k)) - Q(\mathcal{D}, (a_{-k}, a_k'))$ for the effect of axis k when the remaining decisions are held at a_{-k}. A deployed pipeline logs one record $(\phi(\mathcal{D}), a, Q(\mathcal{D}, a))$ per completed build, a single scalar for a single arm. Identifying one Δ_k needs two records that agree on the remaining $K - 1$ coordinates. Uncontrolled logs contain such a pair only by chance, because the space $\mathcal{A} = \prod_k \mathcal{A}_k$ is exponential in K. Without matched records, per-axis effects are not merely noisy but unidentified.

Paired arms. We construct the identifying pair instead of waiting for it. A sweep on axis k fixes a_{-k} and executes builds over $\tilde{\mathcal{A}}_k \subseteq \mathcal{A}_k$, so every ordered pair delivers $m_k = \Delta_k(\mathcal{D}; a_k^+, a_k^-)$ by direct measurement. The margin is the summand that axis k contributes to Eq. 1, hence a policy trained on these pairs is trained on the quantity by which it is judged. Sweeping the model family on AirCraft yields $m = 0.348\ R^2$ for the plain backbone over the geometry-aware one.

Process rewards. Repeated seeds give a deviation $\hat{\sigma}_k$ of Q, from which we set $\tau_k = c\,\hat{\sigma}_k$. A pair is labelled $a_k^+ \succ a_k^-$ when the margin clears τ_k and is retained as a tie otherwise. A tie is evidence against any rule claiming that axis k matters at $\phi(\mathcal{D})$ and is never discarded. Some axes never separate, because the difference between their options is one of correctness. Examples are a split that leaks geometries across folds, and an objective built on an untrained coefficient head, whose gradient is identically zero whatever the accuracy. Such an axis carries a predicate c_k that a single run verifies. The per-decision reward therefore reads the margin on outcome axes and the predicate on validity axes, and abstains otherwise (Appendix B.3, Eqs. 8 and 9). Inverse axes use the same construction, with Q read on the attained objective in absolute terms (Appendix B.3). A sweep therefore returns a preference corpus $\mathcal{C}$ and a tie archive $\mathcal{T}$, scored per decision, and explicit about the decisions it cannot score. Table 3 in Appendix B.2 lists the axes of a build with their options and scoring basis.

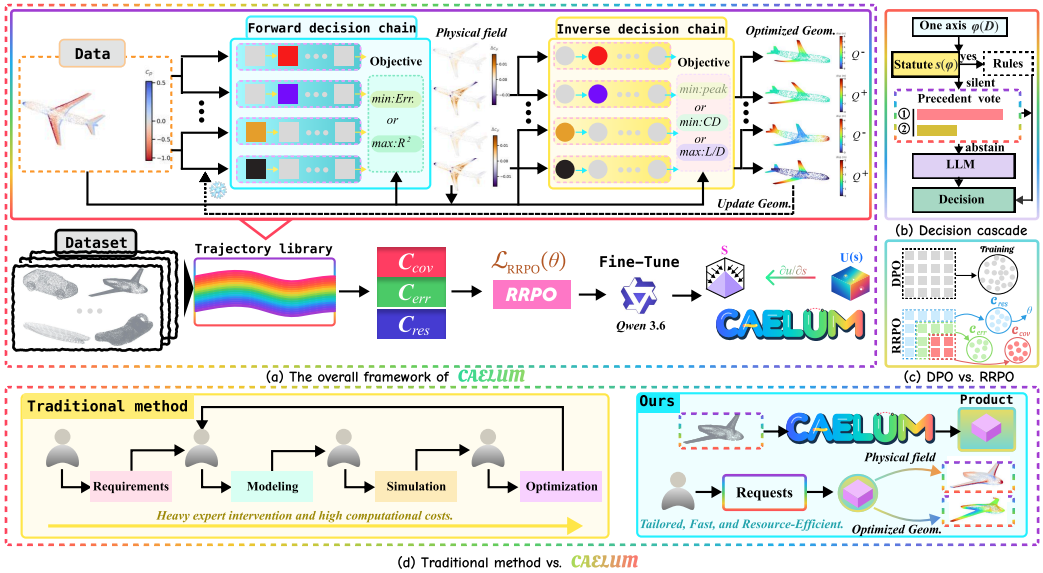
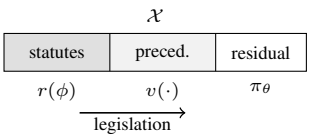


Figure 2: The Caelum framework. (a) Supervision is manufactured by sweeping a dataset over the axes of a build. Arms differ at a single decision (coloured cell), the forward and inverse chains score each arm, and the higher-scoring arm Q^+ is preferred over its counterfactual Q^-. Sweeps accumulate as the corpus $\mathcal{C}$, whose partition feeds RRPO. (b) One decision through the cascade: a statute that fires settles the axis, on silence the precedents vote, and on abstention the open-book weight tier answers. (c) Monolithic DPO trains on the whole corpus, whereas RRPO trains only on $\mathcal{C}_{\text{res}}$. (d) The conventional workflow hand-builds modelling, simulation, and optimization per design. Caelum turns a dataset and a request into a tailored optimizer.

Sweep economy. The design compounds. Axes share a common reference configuration, so covering K axes with m_k options costs $1 + \sum_k (m_k - 1)$ builds and yields $\sum_k \binom{m_k}{2}$ pairs. The evidence grows quadratically in a budget that grows linearly.

4.2 Case-Law Consolidation



A sweep returns $\mathcal{C}$ and $\mathcal{T}$, and the question is where to keep them. Their pairs differ in how well the evidence covers the descriptor space. Some lessons hold under a common predicate and replicate across datasets. Some exist at a single point of $\mathcal{X}$. Some regions carry no evidence at all. Caelum therefore assigns each lesson to a medium according to coverage. A statute answers where one fires. Retrieval answers where precedents are decisive. The weights answer the remainder, and legislation steadily shrinks that remainder. That is how the compiler evolves. Each dataset compiled adds statutes and precedents, and the residual left to the weights narrows.

Statutes. A candidate statute is a pair (p, a) of an interpretable predicate $p : \mathcal{X} \to \{0, 1\}$ and an answer on one axis. It is promoted under three conditions (Appendix B.3, Eq. 10). Its provenance-weighted agreement over the pairs under p must clear a threshold α, where ties count against it, so an axis that fails to separate cannot be legislated. Its support must span at least two datasets. And a must not already dominate outside p. On our corpus, the candidate that assigns the plain backbone to condition-driven datasets is promoted at agreement 1.00 over three datasets. The candidate that assigns it to small-n datasets is rejected at 0.73, because two ties count against it. A promoted statute enters the book $\mathcal{S}_k$ and is evaluated exactly. Insertion perturbs no other entry, and the book grows monotonically with $\mathcal{C}$. A lesson that held on every dataset swept needs no predicate. It enters the book as the updated default of its axis, and the gate legislates the conditional lessons (Table 5).

Precedent retrieval. Evidence at isolated points supports no region-level claim, so the instance is kept verbatim and transferred by similarity. The pairs on an axis cast kernel-weighted votes $v(a)$ (Appendix B.3, Eq. 11). The retrieval tier answers $\hat{a}_{\text{ret}} = \arg\max_a v(a)$ only when the leader clears the runner-up by a factor γ and an absolute floor $v_{\min}$ (Appendix B.3), and abstains otherwise. A new precedent adds a summand and rewrites nothing.

Cascaded arbitration. The three media compose into the policy of Eq. 1,

$$\pi(\phi)_k = \begin{cases} a, & \exists\,(p,a) \in \mathcal{S}_k : \; p(\phi) = 1, \\ \hat{a}_{\text{ret}}, & v(\hat{a}_{\text{ret}}) \;\geq\; \gamma\, v(\hat{a}_2) \;\wedge\; v(\hat{a}_{\text{ret}}) \;\geq\; v_{\min}, \\ \pi_\theta\big(\phi, \mathcal{R}_n(\phi)\big), & \text{otherwise}, \end{cases} \tag{2}$$

where $\hat{a}_2$ is the runner-up, $v_{\min}$ the absolute floor on the leading vote, and $\mathcal{R}_n(\phi)$ the n nearest precedents supplied as context, so the weight tier is never queried bare. Each tier answers only what the previous one cannot, so the weights are consulted only inside the residual that Sec. 4.3 tunes them on.

4.3 Rule-Residual Preference Optimization

The cascade fixes where the weights are consulted. It leaves open what they should be trained on.

Monolithic tuning. The standard answer is to tune the model on everything the agent has learned. Write $\mathcal{L}_{\text{DPO}}(\theta; \mathcal{B})$ for the DPO objective (Rafailov et al., 2023) over a pair set $\mathcal{B}$ against a reference policy π_{ref} (Appendix A, Eq. 4). Monolithic preference tuning is then $\mathcal{L}_{\text{DPO}}(\theta; \mathcal{C})$, the objective taken over the whole corpus (Fig. 2c). It fails at CAE scale. Across three controlled generations of π_θ, every repair broke a lesson that had already been mastered (Fig. 6a), and replaying pairs relocated the failure rather than removing it. Two structural causes compound. First, $|\mathcal{C}|$ stays in the hundreds, so lesson clusters contend for the same weights. Second, most of $\mathcal{C}$ is already answered by $\mathcal{S}$. A gradient step on such a pair moves π_θ away from π_{ref} yet changes no deployed decision at that descriptor.

Residual corpus. We therefore partition the corpus by statute coverage. Writing $s(\phi)$ for the answer the statute tier returns when asked to choose between the pair's two options, undefined when it abstains,

$$\mathcal{C} = \underbrace{\{\, s(\phi) = a^+ \,\}}_{\mathcal{C}_{\text{cov}}} \sqcup \underbrace{\{\, s(\phi) = a^- \,\}}_{\mathcal{C}_{\text{err}}} \sqcup \underbrace{\{\, s(\phi) \text{ undefined} \,\}}_{\mathcal{C}_{\text{res}}}. \tag{3}$$

Pairs in $\mathcal{C}_{\text{cov}}$ are dropped, conflicted pairs in $\mathcal{C}_{\text{err}}$ are routed to adjudication, and only $\mathcal{C}_{\text{res}}$ reaches the weights. Of the 191 usable pairs (Sec. 4.4), 22 are withheld for weak margin or provenance. The statute book covers 137 of the remaining 169, conflicts with none, and leaves 32 residual. With seven degraded-regime control pairs, the residual forms the 39-pair RRPO corpus of Sec. 5.2.

Preference objective. Rule-residual preference optimization is the DPO objective restricted to that support, $\mathcal{L}_{\text{RRPO}}(\theta) = \mathcal{L}_{\text{DPO}}(\theta; \mathcal{C}_{\text{res}})$. The gradient is now supported only where the statute tier abstains, and the tier is tuned in the same open-book regime in which Eq. 2 queries it. The departure from π_{ref} that tuning can induce shrinks with the support. Restricting the support thus avoids paying, for decisions the rules already own, the off-policy distance that drives forgetting (Chen et al., 2026a; Zhang et al., 2026a). Appendix A shows that the monolithic gradient splits exactly into the RRPO gradient and a drift-only component supported where a statute fires, and that the two objectives coincide when $\mathcal{S} = \emptyset$.

4.4 Implementation and Cost

Campaign scale. Each dataset is compiled by one pass of Algorithm 1 (Appendix B). Compiling the fifteen library datasets, about 30,000 solver-computed samples over some 24,000 distinct geometries (Sec. 5, Table 2), logged 601 build-and-evaluate trajectories, and the three aerodynamic datasets added last were each swept over a full twelve-cell build grid at two seeds, with a 30-geometry subset of HiLift swept alongside, 96 builds in all. These distilled into 229 preference records. The generation that tuned the deployed adapter drew on the first 213, of which 191 are usable, and the 16 records logged since then entered the precedent library directly. The legislation gate promoted seven statutes and declined nine.

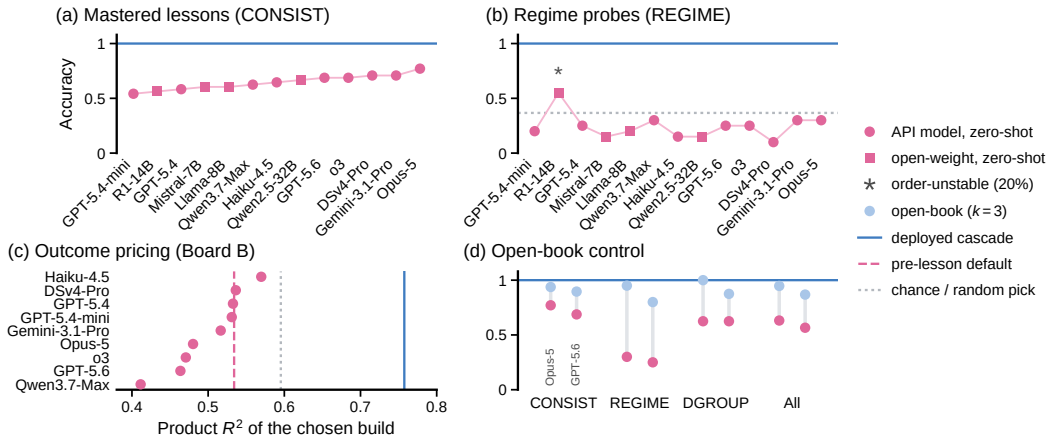


Figure 3: Decision accuracy and its product cost. Pink marks policies without the campaign's evidence, blue the cascade and the arms given its precedents. (a, b) Board A accuracy of thirteen zero-shot models on mastered lessons and regime probes. (c) Board B product R^2 of the builds each model's decisions select. (d) The two strongest arms rerun open-book with the weight tier's retrieved precedents.

Arbiter. The first two tiers carry no learned parameters and run on CPU in seconds. The third is Qwen3.6-35B-A3B, an open-weight 35B mixture-of-experts instruction model adapted by LoRA (Hu et al., 2022). One policy update takes about twenty minutes on four GPUs. The campaign trained twelve such adapters across policy generations and seed replicates. The build system and the cost asymmetry that keeps the campaign affordable are given in Appendix B.

5 Experiments

We evaluate Caelum on fifteen CAE datasets from four physics domains, at two levels: the build decisions a policy makes and the products those decisions ship.

Datasets. As summarized in Table 2 (Appendix B.1), the library holds fifteen datasets. For aerodynamics, these are DrivAerNet++ (Elrefaie et al., 2024), BlendedNet (Sung et al., 2025b), BlendedNet++ (Sung et al., 2025a), NASA CRM (Vassberg et al., 2008), AirCraft (Luo et al., 2025), our synthetic AF-Panel, SuperWing (Yang et al., 2026), HiLift (Ashton et al., 2026), and AeroCar (Liu et al., 2026b;c). For hydrodynamics, they are DTC Hull (el Moctar et al., 2012) and our SUBOFF campaign on the DARPA SUBOFF hull (Groves et al., 1989). For structural analysis, they are SimJEB (Whalen et al., 2021) and DeepJEB (Hong et al., 2025), and for crash dynamics, Car-Crash (Wu et al., 2026).

Evaluation boards. We score decisions on two boards. Board A holds 38 items, each asking which option to take on one build axis for a described dataset: 24 lessons the statute book answers (CONSIST), 10 probes of the data regime (REGIME), and 4 statute-silent cases that only precedents can answer (DGROUP). Each item is scored under both option orders, 76 rows in all, with prompt and language controls in Appendix B.4. Board B holds seven items whose candidate builds we trained and cached, so a policy is scored by the held-out R^2 of the build its answer selects, and regret is the gap to the oracle in R^2 points, the single-axis case of Eq. 1. Both boards are written from the campaign that the cascade consolidates, so the cascade is their ceiling rather than a competitor.

Baselines. Following the backbone-scan convention of MLE-bench (Chan et al., 2025), we compare nine frontier API models and four open-weight models with 7B to 32B parameters. All models receive the same prompt and answer zero-shot. As references we report three policies that call no language model: the statutes alone, the precedent vote alone, and the deployed cascade. Board B further includes the per-item oracle, the expected score of a uniformly random pick, and the default policy that a build follows before any sweep.

Table 1: Board B in product R^2. Each policy is scored by the held-out R^2 of the builds its decisions select, under the deployed prompt (DEPL.) and over all 56 frozen rows (ALL). Regret is the deployed gap to the oracle in R^2 points. PROMOTION: margin of the cascade over the best zero-shot model (underlined).

POLICY	EXAM	PRODUCT R^2		REGRET ↓
		DEPL. ↑	ALL ↑	
ORACLE	/	0.7586		0.00
RANDOM PICK	/	0.5954		16.32
DEFAULT POLICY	/	0.5341		22.45
Haiku-4.5	0.482	<u>0.5697</u>	0.5103	18.89
DSv4-Pro	0.339	0.5363	0.5070	22.23
GPT-5.4	0.464	0.5326	<u>0.5118</u>	22.59
GPT-5.4-mini	0.446	0.5310	0.5002	22.76
Gemini-3.1-Pro	0.482	0.5165	0.5005	24.21
Opus-5	0.464	0.4802	0.4980	27.84
o3	0.464	0.4707	0.4702	28.79
GPT-5.6	0.357	0.4636	0.4643	29.49
Qwen3.7-Max	0.339	0.4114	0.4309	34.72
DEPLOYED CASCADE	**1.000**	**0.7575**		**0.11**
PROMOTION	/	+0.188	+0.246	/

5.1 MAIN RESULTS

Zero-shot comparison. As shown in Fig. 3(a, b), the thirteen models behave differently on the two groups of Board A. On the mastered lessons (CONSIST), accuracy increases with model capability, from 0.54 for the weakest model to 0.77 for the strongest. On the regime probes (REGIME), however, every model but one falls below the chance level of 0.37, scoring between 0.10 and 0.30. The exception is order-unstable (asterisk in Fig. 3b, Appendix C). Neither a larger scale nor a different vendor improves the result. In contrast, the deployed cascade answers all regime probes correctly (accuracy 1.00), and on the four statute-silent items the precedent vote clears its mandate every time (Fig. 4a). The gap is one of evidence rather than capability: given the weight tier's own retrieved precedents, the strongest models recover to 0.87–0.95 (Fig. 3d, Appendix C).

Cost of wrong decisions. Board B scores each answer by the held-out R^2 of the build it selects (Fig. 3c, Table 1). The deployed cascade reaches a product R^2 of 0.7575, only 0.11 points below the oracle, and exceeds the best zero-shot model (Haiku-4.5, 0.5697) by 0.188. All nine frontier models fall below the expected score of a uniformly random pick (0.5954, with significance tests in Appendix C). The loss comes mainly from the regime-decisive items: 91% of the frontier votes on these items go to the geometry-aware variants, which collapse on those datasets. How much such a decision costs depends on the regime. On the full build grids of the three datasets added last, the family axis is flat on SuperWing and AeroCar and a cliff on HiLift, where most geometry-aware cells collapse or diverge while every plain cell exceeds 0.92 (Fig. 4b).

5.2 ANALYSIS

Forward and inverse products. Fig. 5 shows both products of SuperWing, HiLift, and AeroCar from the deployed build and from the build in which one decision reverts to the pre-lesson default. On the forward side, the deployed builds reach a per-case R^2 of 0.99, 0.94, and 0.85 on SuperWing, HiLift, and AeroCar. The wrong decision costs little on SuperWing (0.98) and AeroCar (0.81) but leaves HiLift with an almost constant field (R^2 of −0.00). On the inverse side, each wrong decision breaks the design loop in a different way. An uncapped budget lets the SuperWing loop report a 45.7% improvement at 22.9% displacement, far outside the trust region. The pre-lesson surrogate produces NaN on HiLift at the second step, and the detached C_D head freezes the AeroCar objective. The deployed loops improve their surrogate-evaluated objectives by 4.3%, 1.03%, and 25.9% within 1.40%, 0.07%, and 0.85% displacement (Appendix D).

Tuning a second base model. To test whether the tuning recipe depends on the 35B arbiter, we repeat it on Llama-3.1-8B-Instruct with the two corpora of the deployed generation, the monolithic

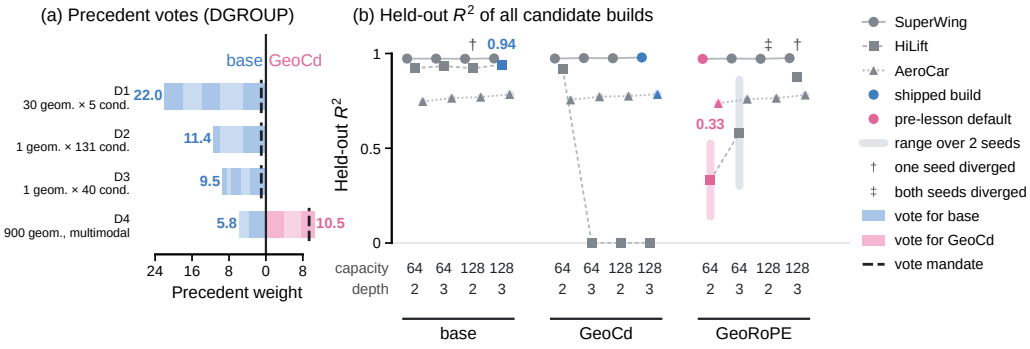


Figure 4: Arbitration where statutes are silent, and what a decision is worth. (a) Kernel-weighted precedent votes on the four statute-silent items of Board A, base to the left and GeoCd to the right, with the mandate each leader must clear. (b) Held-out R^2 of every candidate build on SuperWing, HiLift, and AeroCar, with the shipped cell and the pre-lesson default cell marked.

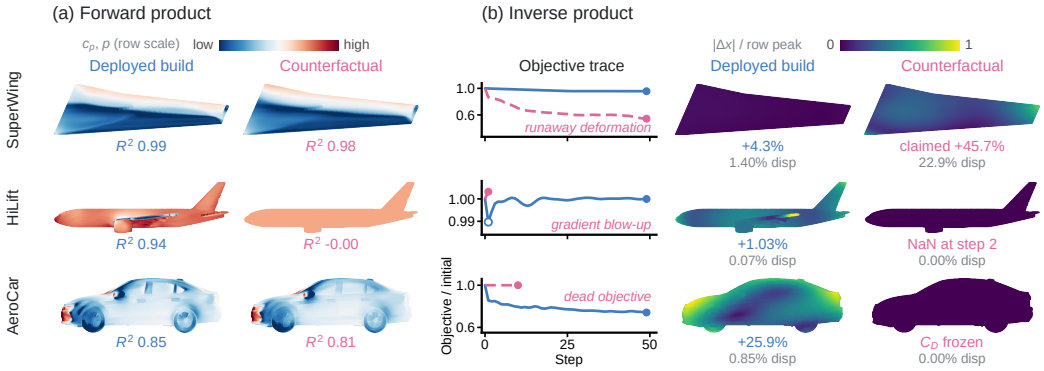


Figure 5: Forward and inverse products of SuperWing, HiLift, and AeroCar, from the deployed build (blue) and from a build with one decision reverted to the pre-lesson default (pink). (a) Predicted surface field on the R^2-median validation case. (b) Objective trace of the design loop and the optimized geometry shaded by displacement.

training set of 206 rows and its 39-pair residual (Sec. 4.3). Both raise exam accuracy above the zero-shot 0.447. The monolithic corpus peaks at 0.684, and the residual corpus reaches 0.697 with 19% of the rows, lifting REGIME accuracy from 0.20 to 0.65 and DGROUP from 0.12 to 0.88 (Table 8, Appendix C). Swapping only the weight-tier adapter inside the deployed cascade leaves every deployed decision unchanged (Table 9), so RRPO buys the weight tier's coverage of statute-silent descriptors with 19% of the corpus and none of the drift on covered descriptors.

6 Conclusion

Caelum turns raw CAE datasets into deployed forward–inverse optimizers and evolves with every dataset it compiles. It is supervised by counterfactual sweeps, consolidated as case law, and tuned only on the rule residual. It compiles fifteen datasets end-to-end, answers the regime probes that frontier models miss, and matches the full-corpus adapter on every deployed decision with 19% of its pairs. Future work couples shipped designs to solver-in-the-loop verification and extends to volumetric and transient physics.

AI USE STATEMENT

We used a large language model (LLM) for language polishing, to improve clarity, for example by correcting grammar and refining phrasing for greater 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.

ETHICS STATEMENT

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

REPRODUCIBILITY STATEMENT

Implementation details, datasets, and evaluation protocols are provided in Appendix B. Full results and forward–inverse evaluations are reported in Appendices C and D.

REFERENCES

Agnar Aamodt and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

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.

Kelsey R Allen, Tatiana Lopez-Guevara, Kimberly Stachenfeld, Alvaro Sanchez-Gonzalez, Peter Battaglia, Jessica B Hamrick, and Tobias Pfaff. Physical design using differentiable learned simulators. *arXiv preprint arXiv:2202.00728*, 2022.

Neil Ashton, Adam Clark, Liam Heidt, Christopher Ivey, Sanjeeb Bose, Rahul Agrawal, Konrad Goc, Rishi Ranade, Corey Adams, Peter Sharpe, Sheel Nidhan, Semit Akkurt, Daniel Leibovici, and Jean Kossaifi. HiLiftAeroML: High-fidelity computational fluid dynamics dataset for high-lift aircraft aerodynamics. *arXiv preprint arXiv:2605.19565*, 2026.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Madry. MLE-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. In *International Conference on Machine Learning*, 2026a.

Jingwen Chen, Wenkai Yang, Shengda Fan, Wenbo Nie, Chenxing Sun, Shaodong Zheng, Yangen Hu, Lu Pan, Ke Zeng, and Yankai Lin. Rethinking continual experience internalization for self-evolving LLM agents. *arXiv preprint arXiv:2606.04703*, 2026b.

Yuxuan Chen, Xu Zhu, Hua Zhou, and Zhuyin Ren. MetaOpenFOAM: an LLM-based multi-agent framework for CFD. *arXiv preprint arXiv:2407.21320*, 2024.

Ould el Moctar, Vladimir Shigunov, and Tobias Zorn. Duisburg test case: Post-panamax container ship for benchmarking. *Ship Technology Research*, 59(3):50–64, 2012.

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, Datasets and Benchmarks Track*, 2024.

E Fan, Kang Hu, Zhuowen Wu, Jiangyang Ge, Jiawei Miao, Yuzhi Zhang, He Sun, Weizong Wang, and Tianhan Zhang. ChatCFD: An LLM-driven agent for end-to-end CFD automation with structured knowledge and reasoning. *arXiv preprint arXiv:2506.02019*, 2025.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural Information Processing Systems*, 2024.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Transactions on Machine Learning Research*, 2026.

Nancy C Groves, Thomas T Huang, and Ming S Chang. Geometric characteristics of DARPA SUBOFF models. Technical Report DTRC/SHD-1298-01, David Taylor Research Center, 1989.

Kostas Hatalis, Despina Christou, and Vyshnavi Kondapalli. Review of case-based reasoning for LLM agents: Theoretical foundations, architectural components, and cognitive integration. *arXiv preprint arXiv:2504.06943*, 2025.

Seongjun Hong, Yongmin Kwon, Dongju Shin, Jangseop Park, and Namwoo Kang. DeepJEB: 3D deep learning-based synthetic jet engine bracket dataset. *Journal of Mechanical Design*, 147(4): 041703, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren (eds.). *Automated machine learning: Methods, systems, challenges*. Springer, 2019.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-driven exploration in the space of code. *arXiv preprint arXiv:2502.13138*, 2025.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

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.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Jianzhe Lin, Fei Wang, Xiaolin Li, Rajeshkumar Golani, and Jubin Chheda. Repeated post-training is not self-improving: Diagnosing scientific amnesia in continual DPO pipelines. *arXiv preprint arXiv:2606.21089*, 2026.

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.

Ye Liu, Shouyi Liu, Huiyu Yang, Jianghang Gu, Wenhao Fan, Zhongxin Yang, Ding Wang, Simeng Chen, Zirun Jiang, Yuanwei Bin, Shiyi Chen, and Yuntian Chen. AeroAgent: A vision–physics–decision framework for aerodynamic vehicle design. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026c.

Darui Lu, Jordan M. Malof, and Willie J. Padilla. An agentic framework for autonomous metamaterial modeling and inverse design. *ACS Photonics*, 12(11):6071–6080, 2025.

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.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Elsevier, 1989.

NVIDIA Corporation. NVIDIA PhysicsNeMo: An open-source framework for physics-based machine learning. https://github.com/NVIDIA/physicsnemo, 2025. Formerly NVIDIA Modulus.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026.

Sandeep Pandey, Ran Xu, Wenkang Wang, and Xu Chu. OpenFOAMGPT: a RAG-augmented LLM agent for OpenFOAM-based computational fluid dynamics. *arXiv preprint arXiv:2501.06327*, 2025.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

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

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Nicholas Sung, Steven Spreizer, Mohamed Elrefaie, Matthew C Jones, and Faez Ahmed. BlendedNet++: A dataset and benchmark for field-resolved aerodynamics and inverse design of blended wing body aircraft. *arXiv preprint arXiv:2512.03280*, 2025a.

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. *arXiv preprint arXiv:2509.07209*, 2025b.

Mingxu Tao, Jiawei Hu, Xian Zhou, Wenpeng Hu, Jiajun Cheng, Yunbo Cao, Zhunchen Luo, and Guotong Geng. When rules learn: A self-evolving agent for legal case retrieval. In *Findings of the Association for Computational Linguistics: ACL*, 2026.

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 2008-6919.

Eamon Whalen, Azariah Beyene, and Caitlin Mueller. SimJEB: Simulated jet engine bracket dataset. *Computer Graphics Forum*, 40(5):9–17, 2021.

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.

Yunjia Yang, Weishao Tang, Mengxin Liu, Nils Thuerey, Yufei Zhang, and Haixin Chen. SuperWing: A comprehensive transonic wing dataset for data-driven aerodynamic design. *Scientific Data*, 13:1113, 2026.

Ling Yue, Nithin Somasekharan, Tingwen Zhang, Yadi Cao, Zhangze Chen, Shimin Di, and Shaowu Pan. Foam-Agent: A large language model-based multi-agent framework for automating computational fluid dynamics workflows. *arXiv preprint arXiv:2505.04997*, 2025.

Miaosen Zhang, Yishan Liu, Shuxia Lin, Xu Yang, Qi Dai, Chong Luo, Weihao Jiang, Peng Hou, Anxiang Zeng, Xin Geng, and Baining Guo. Towards on-policy SFT: Distribution discriminant theory and its applications in LLM training. In *International Conference on Machine Learning*, 2026a.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations*, 2026b.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *AAAI Conference on Artificial Intelligence*, 2024.

Hongyu Zhao, Siyu Zhou, Haolin Yang, Zengyi Qin, and Tianyi Zhou. Neuro-symbolic synergy for world modeling. *arXiv preprint arXiv:2602.10480*, 2026.

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.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, and Jun Wang. Memento: Fine-tuning LLM agents without fine-tuning LLMs. *arXiv preprint arXiv:2508.16153*, 2025.

A Monolithic Gradient Decomposition

This appendix states and derives the decomposition summarized at the end of Sec. 4.3: monolithic preference tuning equals RRPO plus a component that cannot change the deployed policy.

Setup. Recall the DPO objective (Rafailov et al., 2023) over a finite pair set $\mathcal{B}$,

$$\mathcal{L}_{\mathrm{DPO}}(\theta;\mathcal{B}) = -\frac{1}{|\mathcal{B}|}\sum_{(\phi,\, a^+\succ a^-)\in\mathcal{B}} \log\sigma\left(\beta\log\frac{\pi_\theta(a^+\,|\,\phi)}{\pi_{\mathrm{ref}}(a^+\,|\,\phi)} - \beta\log\frac{\pi_\theta(a^-\,|\,\phi)}{\pi_{\mathrm{ref}}(a^-\,|\,\phi)}\right), \tag{4}$$

and the partition of Eq. 3, $\mathcal{C} = \mathcal{C}_{\mathrm{cov}} \sqcup \mathcal{C}_{\mathrm{err}} \sqcup \mathcal{C}_{\mathrm{res}}$. The partition is by whether the statute tier, asked to choose between the pair's two options on its axis, returns a^+, returns a^-, or abstains.

Decomposition. Because Eq. 4 is an average of per-pair terms, splitting the sum over the partition gives, for $\lambda = |\mathcal{C}_{\mathrm{res}}|/|\mathcal{C}|$,

$$\mathcal{L}_{\mathrm{DPO}}(\theta;\mathcal{C}) = \lambda\,\mathcal{L}_{\mathrm{DPO}}\big(\theta;\mathcal{C}_{\mathrm{res}}\big) + (1-\lambda)\,\mathcal{L}_{\mathrm{DPO}}\big(\theta;\mathcal{C}_{\mathrm{cov}}\sqcup\mathcal{C}_{\mathrm{err}}\big), \tag{5}$$

and the same identity holds for the gradients by linearity. The first term is $\lambda\,\mathcal{L}_{\mathrm{RRPO}}(\theta)$ by the definition of Sec. 4.3.

Nullity of the second component. Fix any pair in $\mathcal{C}_{\mathrm{cov}} \sqcup \mathcal{C}_{\mathrm{err}}$ and its descriptor ϕ. By construction a statute fires at ϕ on the pair's axis k and answers with one of the pair's own options, so on the full menu $\mathcal{A}_k$ the first branch of Eq. 2 applies. The deployed decision is therefore

$$\pi(\phi)_k = s(\phi), \tag{6}$$

a quantity that does not depend on θ. Hence for every such ϕ,

$$\frac{\partial}{\partial\theta}\,\pi(\phi)_k = 0\,: \tag{7}$$

no gradient step taken on the second component of Eq. 5 can alter any decision the cascade emits at those descriptors. The component is nevertheless nonzero in general, so following it moves π_θ away from π_{ref}. It pays the distribution shift that drives forgetting (Chen et al., 2026a; Zhang et al., 2026a) while purchasing no deployed decision at the covered descriptors. Through the shared parameters it can still move the answers at uncovered descriptors, in either direction, so the decomposition bounds what the component can buy rather than its value. Appendix C measures that value: the covered pairs raise the bare adapter's closed-book accuracy on statute-silent probes and change no decision of the deployed cascade.

Consequences. Three special cases organize the design space. (i) With an empty statute book, $\mathcal{S} = \emptyset$, no statute fires, so $\mathcal{C}_{\mathrm{cov}} = \mathcal{C}_{\mathrm{err}} = \emptyset$ and $\lambda = 1$. Eq. 5 then collapses to $\mathcal{L}_{\mathrm{DPO}}(\theta;\mathcal{C}) = \mathcal{L}_{\mathrm{RRPO}}(\theta)$, so monolithic tuning is the degenerate case of RRPO without legislation. (ii) As legislation accumulates, λ contracts monotonically, and with it the support on which the weights are trained and the drift the tuning can induce. (iii) Pairs in $\mathcal{C}_{\mathrm{err}}$ are excluded from training in either regime. A conflict between a rule and a pair means that one of them is wrong. The pair is therefore routed to adjudication, which either amends the statute or discards the pair, rather than letting a gradient average the contradiction into the weights. On our corpus the set is empty.

In the deployed generation, the 169 pairs that reach any training set partition into $|\mathcal{C}_{\mathrm{cov}}| = 137$, $|\mathcal{C}_{\mathrm{err}}| = 0$, and $|\mathcal{C}_{\mathrm{res}}| = 32$, giving $\lambda \approx 0.19$. By the argument above, four fifths of the monolithic gradient is drift-only. That is the fraction RRPO declines to pay.

B Implementation Details

This appendix records the configuration of the deployed system and the protocol under which every number in the paper was produced. It covers the benchmarks (B.1), the shipped build of every product (B.2), the arbiter and sweep configuration (B.3), the evaluation protocol (B.4), and the resource account of the study (B.5). Algorithm 1 summarizes one compilation pass end-to-end. Blue marks every point where the decision comes from the arbiter cascade instead of a fixed pipeline.

Build system. The pass adapts a raw dataset to a canonical schema with a by-geometry split and decides the twelve axes of Table 3. It then trains the surrogate on at least two seeds and registers the judged build with its FFD plan and objective.

Algorithm 1 Dataset-to-optimizer compilation in Caelum

Input Data: raw CAE dataset $\mathcal{D}_{\text{raw}}$; objective request o
Input Config: decision axes $\{A_k\}_{k=1}^{K}$ (Table 3), seed budget $S \geq 2$, statute book $\mathcal{S}$, precedent library $\mathcal{P}$, policy π_θ
Output: deployed forward–inverse optimizer Π; updated case law $(\mathcal{S}, \mathcal{P}, \theta)$
// Step 1: Ingest and describe
$\mathcal{D} \leftarrow \text{Adapt}(\mathcal{D}_{\text{raw}})$ // Canonical schema; by-geometry split.
$\phi \leftarrow \text{Descriptors}(\mathcal{D})$ // Scale, physics kind, modality, geometry/condition ratio.
Blue symbols indicate decisions emitted by the arbiter cascade of Eq. 2.
// Step 2: Decide the build plan
for A_k **in** $\{A_1, \ldots, A_K\}$ **do**
  if a statute in $\mathcal{S}$ fires at ϕ on A_k **then** $a_k \leftarrow s(\phi)$ // Rule tier.
  else if the precedent vote clears the γ mandate of Eq. 11 **then** $a_k \leftarrow \hat{a}_{\text{ret}}(\phi)$ // Precedent tier.
  else $a_k \leftarrow \pi_\theta(\phi, A_k)$ // Weight tier, tuned only on the residual (Sec. 4.3).
  end if
end for
// Step 3: Build, judge, and ship
for s **in** $\{1, \ldots, S\}$ **do**
  $\Pi^s \leftarrow \text{Build}(\mathcal{D}, \{a_k\})$ // Train the forward surrogate under the decided plan.
  $r^s \leftarrow \text{Judge}(\Pi^s)$ // Pooled held-out field R^2 (Appendix B.4).
end for
$\Pi \leftarrow$ the judged build, registered with its FFD plan and objective o // Forward + inverse.
// Step 4: Manufacture supervision by counterfactual sweeps
for each paired arm (a, a') on an axis A_k **do**
  build both under the identical protocol; $m \leftarrow r(a) - r(a')$ // Sec. 4.1.
  if $|m| > \hat{\sigma}_k$ **then** add the process-rewarded pair $(\phi,\ a \succ a',\ g)$ to $\mathcal{C}$ // One seed deviation.
  end if
end for
// Step 5: Consolidate into case law and tune the residual
$\mathcal{S} \leftarrow \mathcal{S} \cup \text{Legislate}(\mathcal{C})$ // Promotion gate of Eq. 10; conflicts go to adjudication.
$\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{C}$; $\theta \leftarrow \text{RRPO}(\mathcal{C}_{\text{res}})$ // Weights absorb only what no rule answers (Eq. 3).
Return Π, $(\mathcal{S}, \mathcal{P}, \theta)$

B.1 Benchmarks

Aerodynamics. DrivAerNet++ pairs 7673 production-car variants with a single operating point each (a purely geometry-driven regime) and supervises static pressure and wall shear stress. BlendedNet spans 1099 blended-wing-body planforms under nine flight conditions with four-dimensional conditioning (Sung et al., 2025b), and BlendedNet++ scales the family to 9696 field-resolved geometries (Sung et al., 2025a). NASA CRM is the opposite extreme: one transonic airframe under 40 operating conditions, the condition-driven regime in which the statute selects the plain backbone. AirCraft holds 30 airframes under five conditions and supervises six aerodynamic coefficient component fields. It is the dataset of the $m = 0.348$ example of Sec. 4.1. It is also the dataset on which the untrained-coefficient-head validity predicate retired a dead inverse objective in favour of the peak-field objective it now ships. AF-Panel contributes two synthetic airfoil sets (thin and cambered): NACA four-digit sections labelled by a two-dimensional panel-method solver, supervising c_p on sections of 158 surface points. They are the smallest compilations in the library. SuperWing supervises c_p and c_f on 268 wing geometries under seven flight conditions, and HiLift CRM-HL supervises pressure and wall shear stress on 180 high-lift configurations under ten.

Hydrodynamics. DTC Hull supervises free-surface fields (dynamic pressure and water fraction) on 130 container-ship hull variants. SUBOFF is our own dataset: an in-house simulation campaign over 270 parametric variants of the DARPA SUBOFF hull (Groves et al., 1989), supervising c_p, with a mirror-symmetric deformation plan on the inverse side.

Structures. SimJEB supervises five channels (three displacement components, their magnitude, and stress) on 360 jet-engine brackets under four load cases, with stress trained under a log transform. DeepJEB scales the same task to 2138 generatively designed brackets. There the signed stress

Table 2: Datasets compiled in this work, grouped by aerodynamics, hydrodynamics, structures, and crash dynamics. The ratio of distinct geometries to conditions per geometry is the regime cue a statute reads, and #Train counts training samples after the by-geometry split. AF-Panel spans two synthetic airfoil variants.

DOMAIN	DATASET	#GEOM	#COND	#TRAIN	FIELD
AERO	DrivAerNet++ (Elrefaie et al., 2024)	7673	1	6522	p, WSS
	BlendedNet (Sung et al., 2025b)	1099	9	7963	c_p, c_f
	BlendedNet++ (Sung et al., 2025a)	9696	1	8724	c_p, c_f
	NASA CRM (Vassberg et al., 2008)	1	40	40	c_p
	AirCraft (Luo et al., 2025)	30	5	130	c_{ax}
	AF-Panel (synthetic, panel method)	204	1	183	c_p
	SuperWing (Yang et al., 2026)	268	7	1614	c_p, c_f
	HiLift CRM-HL (Ashton et al., 2026)	180	10	1620	p, WSS
	AeroCar (Liu et al., 2026b;c)	1796	1	1616	p, $\mathbf{u}$
HYDRO	DTC Hull (el Moctar et al., 2012)	130	1	110	p_{rgh}
	SUBOFF (in-house)	270	1	270	c_p
STRUCT.	SimJEB (Whalen et al., 2021)	360	4	360	disp., stress
	DeepJEB (Hong et al., 2025)	2138	4	1200	disp., stress
CRASH	Car-Crash (Wu et al., 2026)	1	131	111	σ_{vM}

Table 3: The build decision space. Each axis is scored on an outcome basis or a validity basis.

STAGE	AXIS k	OPTIONS $\mathcal{A}_k$	BASIS
FORWARD MAP	task route	field / param. / gen.	validity
	model family	plain / geo / geo+RoPE	outcome
	capacity	64 / 128	outcome
	learning rate	10^{-3} / 10^{-4}	outcome
	epochs	fixed / converged	outcome
	conditioning	flight / none	outcome
	initialization	warm / scratch	outcome
	split	random / by-geometry	validity
	field transform	log / none	validity
INVERSE LOOP	deform path	FFD / generative	outcome
	objective	peak field / coefficient	validity
	constraints	smooth+sym. / none	validity

channel trains without the log transform, so the field-transform axis is decided per dataset and not per domain.

Crash. Car-Crash pairs one vehicle with 131 impact conditions and supervises the von-Mises stress of the final time step. It is the second condition-driven compilation, and the second on which the statute's plain-backbone answer ships.

AeroCar. AeroCar is the vehicle-aerodynamics corpus of generatively varied car bodies built for UniDrag and AeroAgent (Liu et al., 2026b;c). The source works do not name the dataset, so the label is ours.

B.2 SHIPPED BUILD CONFIGURATIONS

Table 4 records the configuration each library product actually shipped with along the axes of Table 3: the model family and size the policy selected, the conditioning dimension read from the dataset, the supervised channels, and the registered inverse objective with its deformation symmetry. Three regularities are visible at a glance. All three condition-driven datasets (CRM, AirCraft, Car-Crash) ship the plain backbone, which is the condition-driven statute in action. Every inverse loop ships with smoothness constraints on, and all but Car-Crash carry a mirror symmetry. The two

Table 4: Shipped build configurations of the fifteen library products. $h \times L$ is hidden width $\times$ depth, COND. the operating-condition dimension (0: no conditioning), and SYM. the mirror symmetry of the FFD plan (smoothness constraints are on for every product).

	FORWARD BUILD				INVERSE BUILD		HELD-OUT	
PRODUCT	FAMILY	$h \times L$	COND.	FIELDS	OBJECTIVE	SYM.	$R^2 \uparrow$	n
DrivAerNet++	geo+RoPE	128×3	0	4	min C_D	y	0.93	20
BlendedNet	geo+RoPE	64×2	4	4	min C_D	y	0.86	40
BlendedNet++	geo+RoPE	128×3	4	4	min C_D	y	0.89	20
NASA CRM	plain	128×3	6	4	min peak c_p	y	0.89	4
AirCraft	plain	128×3	3	6	min peak c_{ax}	z	0.97	15
AF-Panel (thin)	plain	128×4	1	1	min peak c_p	y, z	0.91	11
AF-Panel (camb.)	geo+RoPE	64×2	1	1	min peak c_p	y, z	0.86	10
DTC Hull	geo+RoPE	128×3	1	2	min peak p_{rgh}	y	0.97	13
SUBOFF	geo+RoPE	64×2	1	1	min peak c_p	y	0.99	30
SimJEB	geo	64×2	4	5	min peak log stress	y	0.58	40
DeepJEB	geo+RoPE	64×2	4	5	min peak stress	y	0.91	40
Car-Crash	plain	128×3	4	1	min peak σ_{vM}	/	0.64	13
SuperWing	geo	128×3	/	/	min peak c_p	/	0.98†	/
HiLift	plain	128×3	/	/	min peak $\lvert\tau\rvert$	/	0.94†	/
AeroCar	geo	128×3	/	/	min peak p	/	0.78†	/

R^2 is the median per-case held-out R^2 over the n validation cases scored for Appendix D. † Scored by the pooled sweep-cell R^2 of the shipped build over two seeds. "/": not applicable or not recorded on this machine.

structural products differ exactly on the field-transform axis (log stress for SimJEB, linear signed stress for DeepJEB), the per-dataset decision of Appendix B.1.

B.3 ARBITER, MODEL ZOO, AND SWEEP PROTOCOL

Process rewards. With margin m_k and threshold $\tau_k = c\,\hat{\sigma}_k$ (Sec. 4.1), each pair is labelled

$$\ell_k = \begin{cases} a_k^+ \succ a_k^-, & m_k > \tau_k, \\ a_k^+ \sim a_k^-, & |m_k| \le \tau_k, \end{cases} \tag{8}$$

and the per-decision reward is

$$r_k = \begin{cases} \rho(m_k), & m_k > \tau_k \quad \text{(outcome)}, \\ \mathbf{1}[\,c_k\,], & c_k \text{ available} \quad \text{(validity)}, \\ \bot, & \text{otherwise}, \end{cases} \tag{9}$$

with ρ monotone.

Legislation and retrieval. Let each pair carry a provenance weight $g_i \in (0, 1]$ set by the protocol that produced it. For a candidate statute (p, a) on axis k, collect the pairs under p that prefer a ($\mathcal{C}^+_{p,a}$), the pairs that reject a ($\mathcal{C}^-_{p,a}$), and the ties that involve a ($\mathcal{T}_{p,a}$). The agreement is

$$\mathrm{agree}(p, a) = \frac{\sum_{i \in \mathcal{C}^+_{p,a}} g_i}{\sum_{i \in \mathcal{C}^+_{p,a}} g_i + \sum_{i \in \mathcal{C}^-_{p,a}} g_i + \sum_{i \in \mathcal{T}_{p,a}} g_i}, \tag{10}$$

and promotion follows the conditions of Sec. 4.2. With a kernel κ on descriptors, the pairs on axis k vote

$$v(a) = \sum_{i \in \mathcal{C}_k} g_i\, \kappa\big(\phi, \phi_i\big)\, \mathbf{1}\big[a_i^+ = a\big], \tag{11}$$

and Eq. 2 mandates the leader only at $v(\hat{a}_{\mathrm{ret}}) \ge \gamma\, v(\hat{a}_2)$ and $v(\hat{a}_{\mathrm{ret}}) \ge 1$ (Table 6). Table 5 lists the gate's output on the campaign corpus. Seven candidate statutes were promoted at unanimous weighted agreement, spanning three axes. Nine further candidates were declined. The gate also backtests each promoted statute against the mastered lessons of Board A and flagged no contradiction.

Table 5: Output of the legislation gate on the campaign corpus: the seven candidate statutes it promoted (all at weighted agreement 1.0), with their evidence base. Nine further candidates were declined by the gate, and no promoted statute contradicted a mastered lesson of Board A.

AXIS	CONDITION	ANSWER	PAIRS	SUPPORTING DATASETS
epochs	geometry driven	converged	3	BlendedNet, DeepJEB
epochs	large n ($\geq$800)	converged	2	BlendedNet, DeepJEB
init mode	small n ($\leq$150)	warm-AB	3	AF-Panel, BlendedNet, CRM
init mode	small + dynamics	warm-AB	3	AF-Panel, BlendedNet, CRM
model family	condition driven	base	5	AirCraft, Car-Crash, CRM
model family	structural multimodal	GeoCd	5	DeepJEB, SimJEB
model family	multimodal	GeoCd	5	DeepJEB, SimJEB

Arbiter. The weight tier π_θ is Qwen3.6-35B-A3B (3B active parameters), adapted by LoRA of rank 16 on the attention projections over a 4-bit quantized base (Hu et al., 2022).

Model zoo. The options on the model-family axis of Table 3 are drawn from an open physics-ML library (NVIDIA Corporation, 2025) that spans solver families for regular grids, meshes, and point clouds. The task route is a validity axis decided before any family is scored, and it eliminates the modalities a dataset cannot use. The three members swept here are the menu that survives for point-cloud surface fields across all fifteen datasets.

Sweep protocol. Every arm is built under the protocol it is meant to inform, with the same schedule, split, and judge as a deployed build. Each arm is repeated over at least two seeds. The judge is pooled field R^2 on the held-out geometries, and we set $\tau_k = \max(\hat{\sigma}_k, 0.01)$ with $\hat{\sigma}_k$ the across-seed sample deviation, so a preference is emitted only when the margin exceeds one seed deviation and an absolute floor of 0.01. Inverse arms are judged on the attained objective in absolute terms, because self-reported relative improvement can mislead. On DeepJEB, a generative path improved the most against its own weak median, yet it reached a peak stress of 549 MPa where the deformation path reached 211 MPa.

Hyperparameters. Table 6 collects every definition and constant of the arbiter as deployed. Because γ and the kernel decide when the retrieval tier answers, the table states them in full.

Release. The frozen boards (`boardA.jsonl` and `boardB.jsonl`, with a manifest carrying their SHA-256 digests), the preference corpus with its tie flags, the residual-corpus partition, the statute book with its declined candidates, the cascade log, the prompt templates of both boards and of the weight tier, and the adapter configurations accompany the paper as supplementary material.

B.4 EVALUATION PROTOCOL AND PROMPT CONTROLS

Parse rule. All arms share one parse rule. Unparseable answers count as wrong and are reported separately (zero on the main protocol).

Prompt controls. Board A items are rendered under two prompt conditions (anonymised and named), two option orders, and two languages (304 rows in a fixed manifest). The deployed protocol is the anonymised Chinese rendering under both option orders (76 rows), and the remaining renderings are controls. Prepending the dataset name moves overall accuracy by at most $+0.02$ on every arm. The only larger per-group shift occurs on items whose names are fictional by construction, where retrieval cannot be the cause. An English-language control moves scores by at most 0.03.

Inverse protocol. Every inverse loop is held inside the surrogate's trust region by construction. Displacements are capped by a per-family budget of a few percent of the feature length. The FFD field carries smoothness and symmetry constraints. The loop ships the tracked best feasible step rather than the final iterate. Reported improvements are therefore claims at those budgets, evaluated by the deployed surrogate. Appendix D re-scores the two products whose solver we own.

Table 6: Arbiter definitions and hyperparameters as deployed. Symbols follow the main text.

QUANTITY	DEFINITION	VALUE
Descriptor and predicates		
$\phi(\mathcal{D})$	Per-dataset record: physics kind, modality, condition-driven flag, dynamics, hard, saturated, zoned, signed stress, n_{train}, number of geometries, cases per geometry	11 fields, matched symbolically
size bucket	Derived from n_{train}	≤ 150 / < 800 / ≥ 800
few-by-many	Derived flag: ≤ 10 geometries with ≥ 5 cases each	binary
p	Hand-written conditions over ϕ (the promoted ones are the conditions of Table 5)	12
Legislation (Eq. 10)		
α	Minimum agreement	0.8
support	Minimum pairs and datasets under p	2 and 2
$\sum g_i$	Minimum provenance weight under p	1.2
contrast	Share of a among the chosen options outside p	$\leq 1/2$
ties	Enter the denominator of Eq. 10 only when they involve a	
backtest	Promoted statute replayed on the mastered lessons of Board A	no contradiction
Provenance weight		
g_i	Production build / curated lesson / full sweep / short sweep or live run / unknown / curated template / cheap protocol	1.0 / 0.9 / 0.8 / 0.7 / 0.6 / 0.5 / 0.35
Retrieval (Eq. 11)		
κ	Sum over fields of a match bonus or a mismatch penalty (weights below)	vote weight $\max(\kappa, 0)\, g_i$
γ	Mandate: leading vote $\geq \gamma\times$ runner-up	1.6
$v_{\min}$	Absolute floor on the leading vote	1.0
k	Precedents in the vote / in the open-book block	5 / 3
Cascade		
statute order	Fixed order, the first firing statute answers, one statute per axis	
alias	Family alias Geo, in builds / on the exam	GeoRoPE / either member
fallback	Answer when every tier abstains	pre-lesson default
Sweep labels		
τ_k	Tie threshold, with $\hat{\sigma}_k$ the across-seed sample deviation of Q	$\max(\hat{\sigma}_k,\ 0.01)$
seeds	Replicates per arm	$\{0, 1\}$
parse	Exact match, then whole-word match, then last non-empty line, else unparseable (wrong)	
Weight tier		
base	Qwen3.6-35B-A3B	4-bit NF4
LoRA	rank / scaling / dropout / target projections	16 / 32 / 0.05 / q, k, v, o
DPO	β / learning rate / epochs	$0.1 / 5 \times 10^{-5}$ / 3

KERNEL κ: FIELD	MATCH	MISMATCH
physics kind	$+2.0$	-1.0
condition-driven flag	$+3.0$	-3.0
few-by-many flag	$+2.0$	-2.0
modality	$+1.5$	-0.5
hard, saturated, dynamics (each)	$+1.0$	-0.5
size bucket	$+1.0$	0
same dataset	$+1.5$	0

B.5 RESOURCE ACCOUNT

The campaign counts of Sec. 4.4 are drawn from the runtime records on disk. Two items complete the account.

Showcase evaluation. The Appendix D galleries evaluate 256 held-out cases and rerun the twelve gallery loops from the production checkpoints. On one GPU of the build node (8× RTX 5090) this replay completes in minutes. The marginal cost of auditing a compiled product is negligible next to the cost of building it.

Cost asymmetry. The construction is affordable because of a cost asymmetry. An arm costs one training run, under two GPU-hours in our largest paired study. One industrial-fidelity aerodynamics label is reported at 6.1×10^4 CPU-hours (Wu et al., 2026). The gap is three to four orders of magnitude.

Table 7: Zero-shot Board A accuracy by group for all thirteen arms, ordered by CONSIST: mastered lessons (CONSIST), regime probes (REGIME), statute-silent cases (DGROUP), and the full board (ALL). Every number is scored on the 76-row deployed protocol.

ARM	ACCESS	CONSIST↑	REGIME↑	DGROUP↑	ALL↑
Opus-5	API	0.771	0.30	0.62	0.632
Gemini-3.1-Pro	API	0.708	0.30	0.62	0.592
DSv4-Pro	API	0.708	0.10	0.50	0.526
o3	API	0.688	0.25	0.38	0.539
GPT-5.6	API	0.688	0.25	0.62	0.566
Qwen2.5-32B	open	0.667	0.15	0.38	0.500
Haiku-4.5	API	0.646	0.15	0.25	0.474
Qwen3.7-Max	API	0.625	0.30	0.50	0.526
Mistral-7B	open	0.604	0.15	0.00	0.421
Llama-8B	open	0.604	0.20	0.12	0.447
GPT-5.4	API	0.583	0.25	0.62	0.500
R1-14B	open	0.562	0.55	0.25	0.526
GPT-5.4-mini	API	0.542	0.20	0.12	0.408

Table 8: Tuning Llama-3.1-8B-Instruct on the monolithic corpus (206 rows) and on the RRPO residual corpus (39 pairs). All configurations share one SFT stage and are scored on the frozen board. Cells with several entries list seeds 42, 7, and 123 in that order. A collapsed seed is marked $\times$ and a seed not run "–". STABLE is the fraction of answers unchanged under option reordering. Bold marks the best cell, underline the best monolithic cell, and PROMOTION the gain of the best cell over zero-shot.

CONFIGURATION			BOARD ACCURACY			ROBUSTNESS	
CORPUS	lr	β	EXAM (BY SEED)↑	REGIME↑	DGROUP↑	UNPARS.↓	STABLE↑
ZERO-SHOT	–	–	0.447	0.20	0.12	0.0%	50%
MONOLITHIC (206)	2e-5	0.1	<u>0.684</u> / – / 0.645	0.60 / 0.50	0.75	1.3%	70 / 70%
	2e-5	0.3	0.632	0.55	0.75	1.3%	60%
	5e-5	0.1	0.632	0.55	0.75	7.9%	90%
	5e-5	0.3	0.684	0.65	0.62	4.0%	90%
	1e-4	0.1	0.658	0.55	0.62	4.0%	90%
	1e-4	0.3	0.671	0.65	0.38	0.0%	40%
RRPO RESIDUAL (39, OURS)	2e-5	0.1	0.592	0.60	0.75	1.3%	80%
	2e-5	0.3	0.539	0.60	0.75	7.9%	100%
	5e-5	0.1	0.632 / 0.645 / $\times$	0.55 / 0.65	0.88	5.3%	50 / 30%
	5e-5	0.3	0.605	0.60	0.88	9.2%	80%
	1e-4	0.1	0.539	0.55	0.75	19.7%	80%
	1e-4	0.3	**0.697** / $\times$ / $\times$	0.65	0.75	2.6%	90%
PROMOTION over zero-shot			+0.250	+0.45	+0.76	/	/

C FULL RESULTS

This appendix completes the numbers behind Sec. 5. Table 7 extends the zero-shot comparison of Fig. 3(a, b) with the statute-silent group (DGROUP) and the overall board accuracy. DGROUP is where precedents, not statutes, must answer. The spread there is wide (0.00–0.62) and only loosely tied to CONSIST rank.

Second-base sweep. Table 8 reports the full sweep on Llama-3.1-8B-Instruct summarized in Sec. 5.2, six configurations per corpus over learning rate and β. The columns beyond the exam separate the corpora on robustness. REGIME rises from 0.20 to 0.55–0.65, and the head configurations stay order-stable at 70–90%. The sweep also exposes a failure mode of its own. At the selected configuration only one seed in three survives. The collapsed seeds degenerate to unparseable text, and their DPO loss stays indistinguishable from healthy runs. The frozen board exposes the collapse and the training loss does not.

Retention generations. Fig. 6a reports the retention gate that preceded RRPO, scored on the 24 CONSIST items. Each monolithic generation retrained a rank-16 adapter on the whole preference corpus after pairs had been added for one further lesson, and every item the previous generation

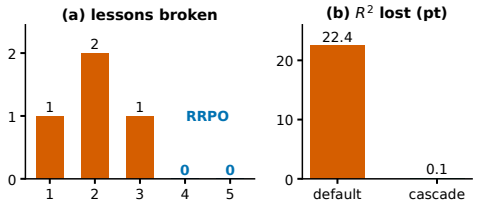


Figure 6: (a) Tuning a bare adapter on the full preference corpus breaks a mastered lesson at every repair. Under case law the same lessons are held by statutes, and the deployed cascade breaks none. (b) Case law cuts the frozen board's decision loss from 22.4 R^2 points under the pre-lesson defaults to 0.1.

Table 9: Weight-tier control on the precedent library of the deployed generation. The cascade is held fixed and only the weight-tier adapter is swapped. Columns count correct answers on the 24 CONSIST items (C), the 10 REGIME probes (R), and the 4 DGROUP items (D), closed-book (bare adapter) and open-book (with the $k = 3$ precedent block the cascade supplies).

	CLOSED-BOOK			OPEN-BOOK		
WEIGHT TIER	C	R	D	C	R	D
Monolithic (206 rows)	21/24	6/10	4/4	24/24	10/10	4/4
Residual (39 pairs, deployed)	19/24	4/10	4/4	24/24	10/10	4/4

answered correctly was then re-scored under greedy decoding. A lesson counts as broken when its item flips from correct to wrong. The first generation added 14 capacity pairs to a corpus of 166 and lost the split lesson. The second grew the corpus to 173 and lost two items, the split lesson among them. The third replayed the lost pairs by oversampling to 206 rows, recovered both, and lost the depth lesson instead. Under RRPO these lessons are held by statutes that retraining cannot touch, so the two residual generations, tuned on 34 and 39 pairs, leave the cascade at 24/24. The comparison is between the bare monolithic adapter and the deployed cascade, because the residual adapter is never queried on a statute-covered item.

Weight-tier control. The retention gate compares a bare adapter with the cascade, so it cannot separate what routing removes from what RRPO adds. Table 9 holds the cascade fixed on the precedent library of the deployed generation and swaps only the weight-tier adapter: the deployed residual adapter (39 pairs) against the monolithic adapter of the same generation (206 rows). On the fourteen statute-silent board items both adapters score 10/10 and 4/4 open-book, as does the precedent vote with no language model at all. The adapters differ only closed-book, where the monolithic one answers more items, the expected return on its larger corpus. On Board B the deployed cascade answers all seven items at the statute and vote tiers, so Table 1 already prices the zero-LLM sub-cascade on the library. The deployed decisions are therefore set by the exact tiers and the retrieved context. The weight tier exists for coverage of statute-silent regions, and RRPO buys that coverage with 19% of the tuning pairs and none of the drift on covered descriptors.

Cascade deficit analysis. The deployed cascade answers every Board B item at the statute tier and is exam-perfect there. Its entire deficit comes from two items where the rule answers the family alias "Geo" and the deployed default resolves it to the wrong member (0.838 against 0.843 and 0.9625 against 0.965). An answer can be exam-correct and still leave R^2 on the table. The dual-level judge is what makes that visible, and the alias itself is now a candidate lesson.

Statistical calibre. Board B holds seven items, so Table 1 differences are read with a paired bootstrap over items (10^4 resamples). The deployed cascade clears the best zero-shot arm (Haiku-4.5) by $+0.188$ with CI_{95} $[+0.044, +0.360]$ and win rate 0.99, and clears the random-pick expectation by $+0.162$, CI_{95} $[+0.063, +0.265]$. All nine frontier point estimates fall below random. Six are individually significant, while the intervals of Haiku-4.5, GPT-5.4-mini, and Opus-5 cross zero. The Board-A groups carry 48/20/8 rows (CONSIST/REGIME/DGROUP), so per-group accuracies are coarse. The dissociation of Fig. 3(a, b) rests on its consistency across thirteen arms, not on any

Table 10: Open-book control on Board A: two frontier arms rerun on the frozen items with the deployed weight tier's own retrieval context appended ($k = 3$). Zero-shot rows are repeated for reference.

ARM	MODE	CONSIST ↑	REGIME ↑	DGROUP ↑	ALL ↑
Opus-5	zero-shot	0.771	0.30	0.62	0.632
Opus-5	open-book	0.938	0.95	1.00	0.947
GPT-5.6	zero-shot	0.688	0.25	0.62	0.566
GPT-5.6	open-book	0.896	0.80	0.88	0.868

single arm. The one REGIME outlier, a 14B distilled model at 0.55, keeps its answers under option reordering only 20% of the time, whereas the strongest models stay order-stable at the wrong answer. Exam accuracy and product R^2 also rank the arms differently: DSv4-Pro has the lowest exam accuracy of the nine but the second-highest product R^2. The largest single-item forfeit is the SimJEB family item, on which four of the nine frontier arms select GeoRoPE, whose build scores R^2 0.193 against 0.683 for the oracle GeoCd.

Open-book control. The zero-shot protocol denies frontier arms the retrieval context that the weight tier receives. We therefore rerun two top arms on the same frozen items with that context appended: the deployed precedent block ($k = 3$, no dataset exclusion, the setting of Eq. 2). Of the 76 rows, 66 retrieve a non-empty block. This arm is diagnostic and sits outside the frozen protocol. Table 10 and Fig. 3d show that the open book carries most of the knowledge: Opus-5 rises from 0.632 to 0.947 overall and from 0.30 to 0.95 on REGIME, GPT-5.6 from 0.566 to 0.868. Neither arm reaches the exact tiers, and both still err on evidence handed to them verbatim. The dissociation of Fig. 3 is therefore a gap in evidence, not in ability.

D PRODUCT SHOWCASES

This appendix renders the library products: the geometry evolution along each optimization trace (Fig. 7), the source datasets (Fig. 8), the shipped forward fields (Fig. 9), and the inverse loops (Fig. 10). SuperWing, HiLift, and AeroCar are rendered in the main text (Fig. 5 and Sec. 5.2), each against a counterfactual row in which one axis reverts to the pre-lesson default. The HiLift objective is the peak wall shear stress. Every row is produced by the corresponding production checkpoint under the protocol of Appendix B.4.

Evolution plates. The evolution plates make the loops' temperament visible. Most rows descend monotonically to a settled plateau: BlendedNet++ reaches $J/J_0 = 0.603$ at 1.06% displacement, DTC Hull 0.760, and DeepJEB 0.841. The remaining rows show why the shipped step is the best feasible step rather than the last one. AirCraft dives to 0.609 by step 15 and climbs back to 0.905 as the learning rate decays, so the open marker ships the early geometry. SUBOFF settles by step 15 (0.934) and never moves again. NASA CRM and Car-Crash move their objectives by at most 0.2%, because the registered peak objective has little leverage on those clouds at this budget.

Field plates. The forward plates are drawn on the R^2-median held-out case of each product, and the error column is normalized per row by the field σ, so rows are comparable across physics. The structural rows are the tail of the library. SimJEB's median case sits at R^2 0.58, with the error concentrated at the bolt bosses, and DeepJEB at 0.91, with residuals along the load-path junctions. The aerodynamic and hydrodynamic rows sit at 0.86–0.99. The dataset plates show four index-spread validation samples per row. They show what one system had to span: four physics domains, geometry families from 158-point airfoil sections to full car bodies, and our own SUBOFF campaign (Appendix B.1).

Inverse plates. The inverse plates print each loop's in-budget improvement next to the displacement that buys it. The two are decoupled: AirCraft reaches −59.5% at 0.08% displacement, BlendedNet++ −41.1% at 1.06%, and DTC Hull −24.8% at 1.38%, whereas DrivAerNet++ moves its full drag objective by −2.9% at 0.65%. The rows that gain nearly nothing (NASA CRM, Car-Crash, and SimJEB, all at or below 0.3%) are the same rows whose traces in Fig. 7 show no leverage. The loop reports the absence of headroom rather than manufacturing a gain.

Table 11: Solver re-evaluation of the two AF-Panel inverse products. J is the loop objective (scale-invariant p-norm of $|c_p|$, $p = 8$) on the initial and the optimized section, read from the deployed surrogate and from the panel solver that labelled the dataset. Displacement is the peak, in fractions of chord.

		Surrogate		Solver			
Product	Case	J_0	ΔJ	J_0	ΔJ	$\max \lvert c_p \rvert$	Disp.
AF-Panel (thin)	NACA 2206, $\alpha = 6^\circ$	2.255	-1.08%	2.520	-1.86%	$4.36 \to 4.27$	1.1%
AF-Panel (camb.)	NACA 4412, $\alpha = 4^\circ$	0.956	-0.69%	1.082	$+0.21\%$	$1.66 \to 1.66$	1.3%

Solver re-evaluation. AF-Panel is the one library dataset whose solver we own, so its two shipped inverse products can be re-scored by the panel method instead of the surrogate (Table 11). The initial and the optimized section of each deployed loop are pushed through the vortex-panel solver that labelled the dataset, at the case's angle of attack, and both are scored with the loop's own objective, the scale-invariant p-norm ($p = 8$) of $|c_p|$. The solver confirms the thin product: the surrogate claims a 1.08% reduction of the peak and the solver measures 1.86%, with the maximum $|c_p|$ falling from 4.36 to 4.27 and lift changed by under 0.5%. It overturns the cambered product: the surrogate claims 0.69% and the solver measures a 0.21% increase, with the maximum $|c_p|$ unchanged to three digits. Both loops moved the section by 1.1 and 1.3% of chord, and on both the surrogate reads the initial peak 10 to 12% low. Inside the trust region the sign of a sub-percent gain is therefore not guaranteed, which is why the solver-in-the-loop judge of Appendix E is the next step. SUBOFF, the other in-house dataset, was labelled by an external CFD campaign whose case setup we do not hold, and the remaining datasets ship no solver.

E Limitations and Future Work

The compiled products are surface-field surrogates and their design loops. Volumetric fields and transient rollouts enter as new task routes without new machinery. One caveat bounds the claims. Inverse improvements are judged by the surrogate that drives the optimization, inside the trust region that the protocol of Appendix B.4 enforces. On the two products we could re-score with their solver, the surrogate's sign held on one and flipped on the other (Table 11). Coupling shipped designs to a solver-in-the-loop judge, affordable because of the cost asymmetry of Appendix B.5, is the immediate next step.

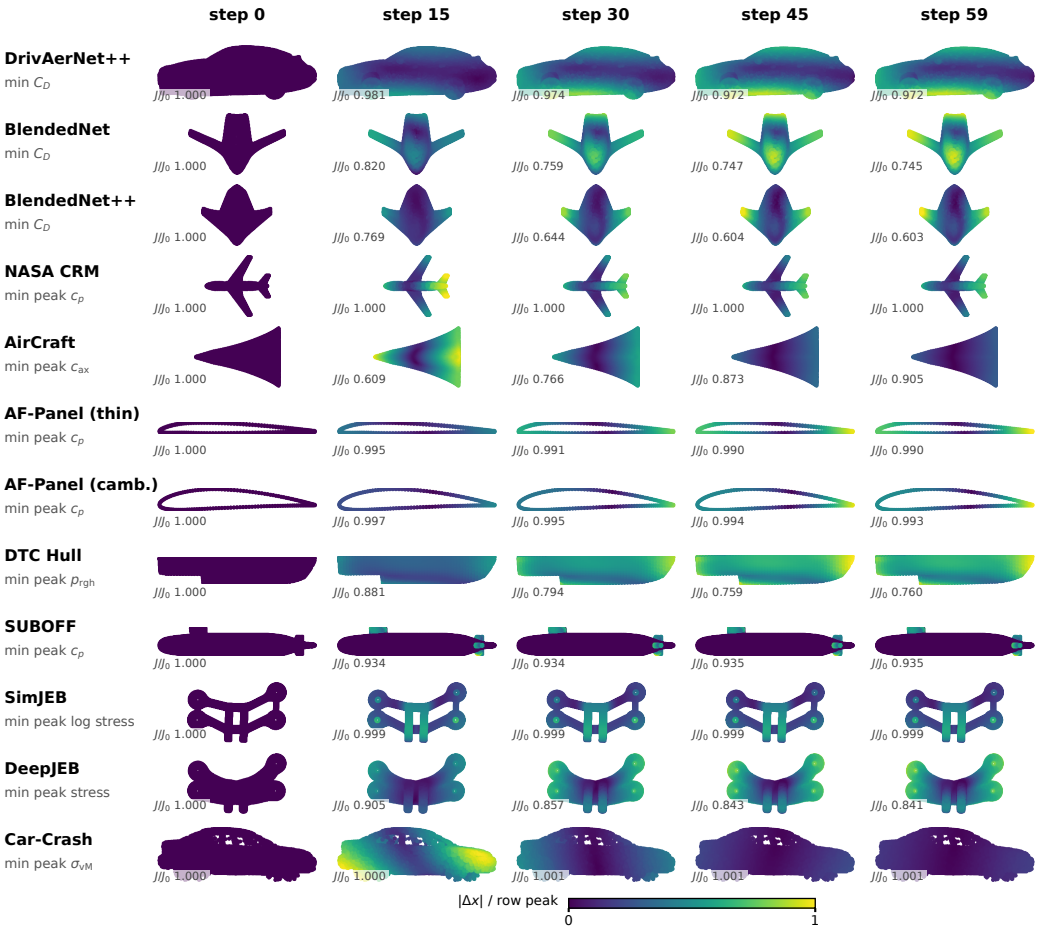


Figure 7: Geometry evolution along the optimization trace of every library product: the running geometry at five snapshots of the inverse loop, shaded by displacement from the initial shape normalized to the row peak (colour bar), with the free-model objective ratio J/J_0 at each snapshot. Monotone rows converge smoothly (BlendedNet++, DTC Hull, DeepJEB), whereas AirCraft overshoots its checkpoint minimum and rebounds.

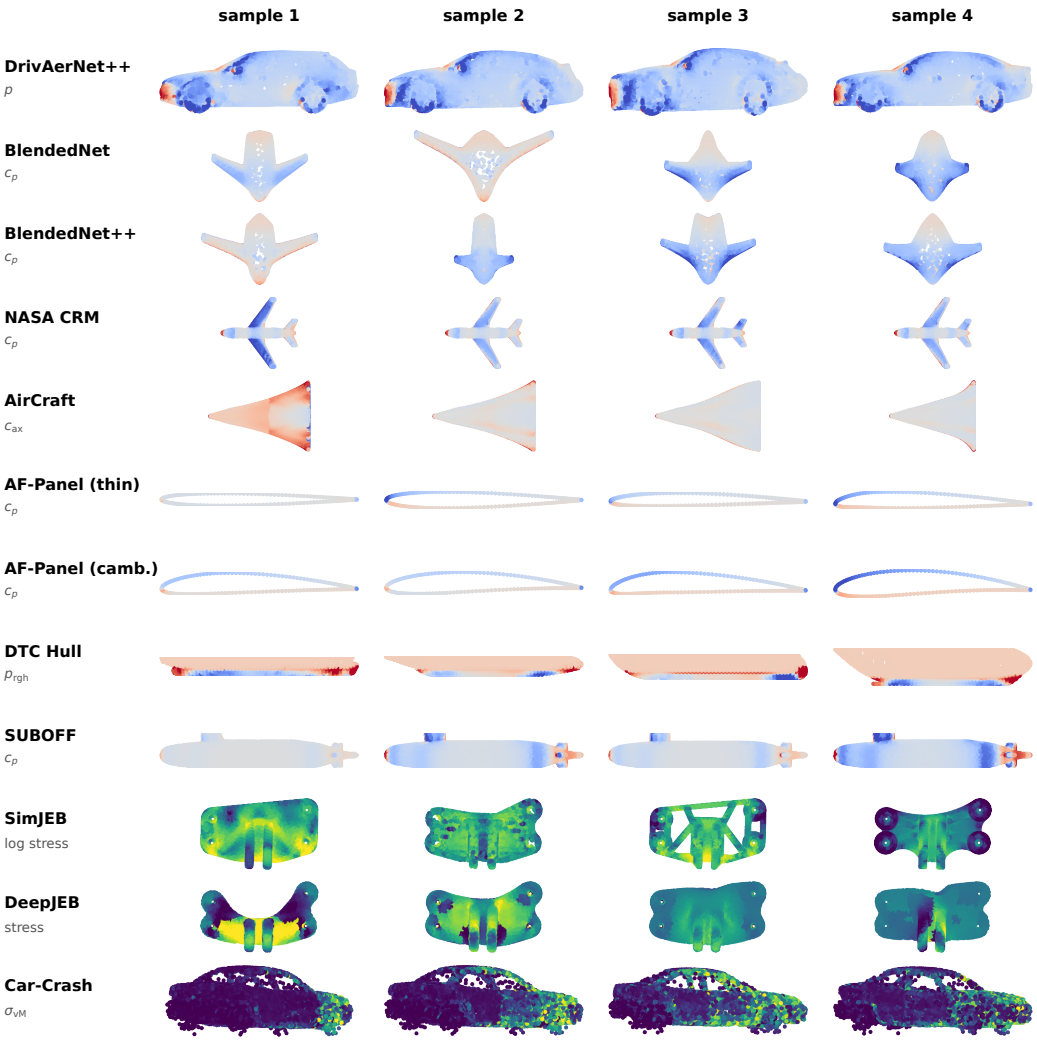


Figure 8: The library datasets. Each row shows four index-spread validation samples of one dataset, coloured by the ground-truth target field on a per-row robust colour scale (diverging for signed fields, sequential otherwise). For NASA CRM and Car-Crash the four samples are four operating conditions of the single geometry (the condition-driven regime that the plain-backbone statute reads). Rows such as BlendedNet, SimJEB, and DeepJEB vary the geometry itself.

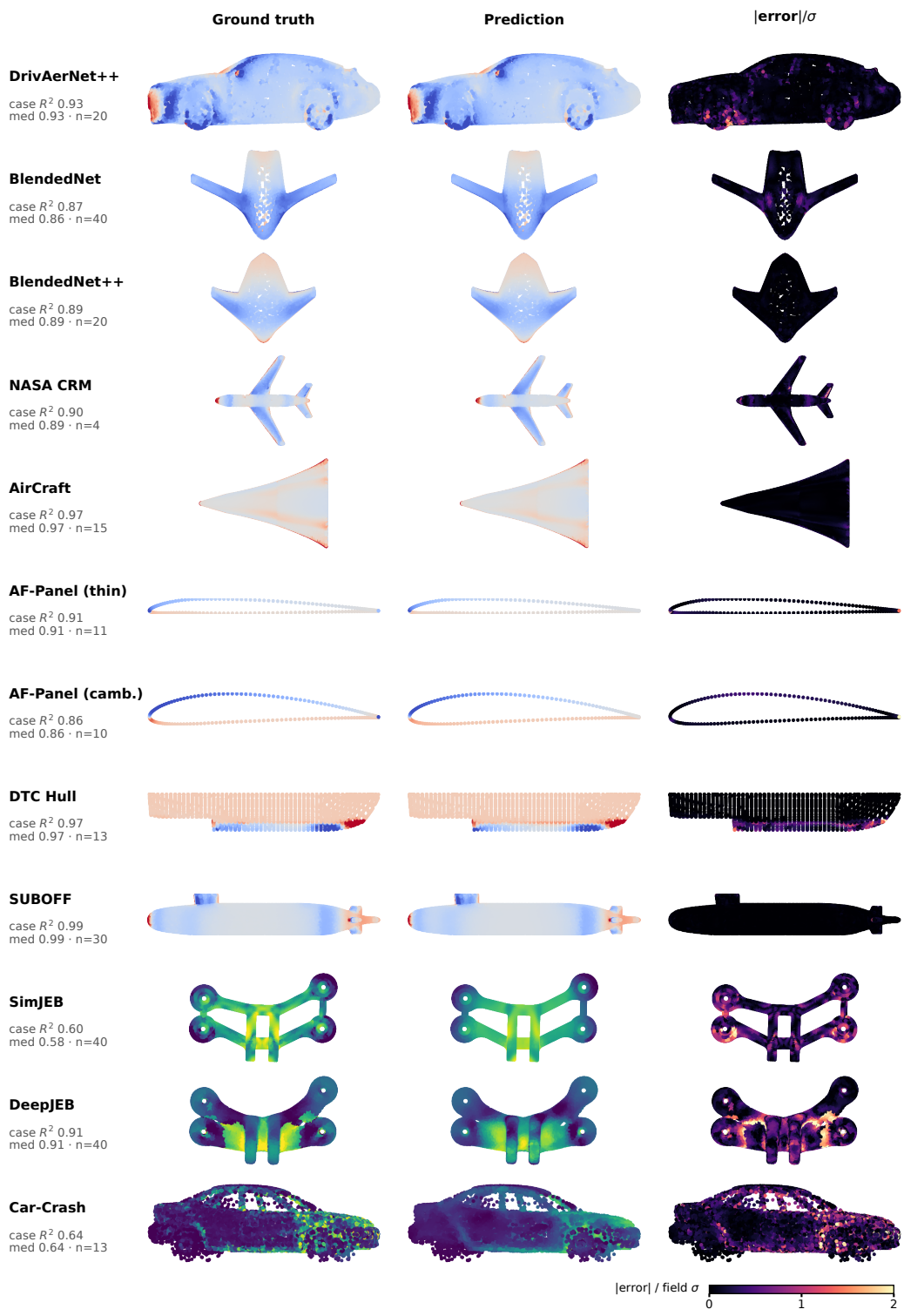


Figure 9: Forward products of the twelve library compilations not shown in the main text. Each row shows the validation case with the median per-case R^2 of the shipped build, a typical case by construction. Columns: ground truth, prediction, and |error| normalized by the field standard deviation (shared scale, colour bar). Badges give the per-case R^2 and the median over the scored validation cases.

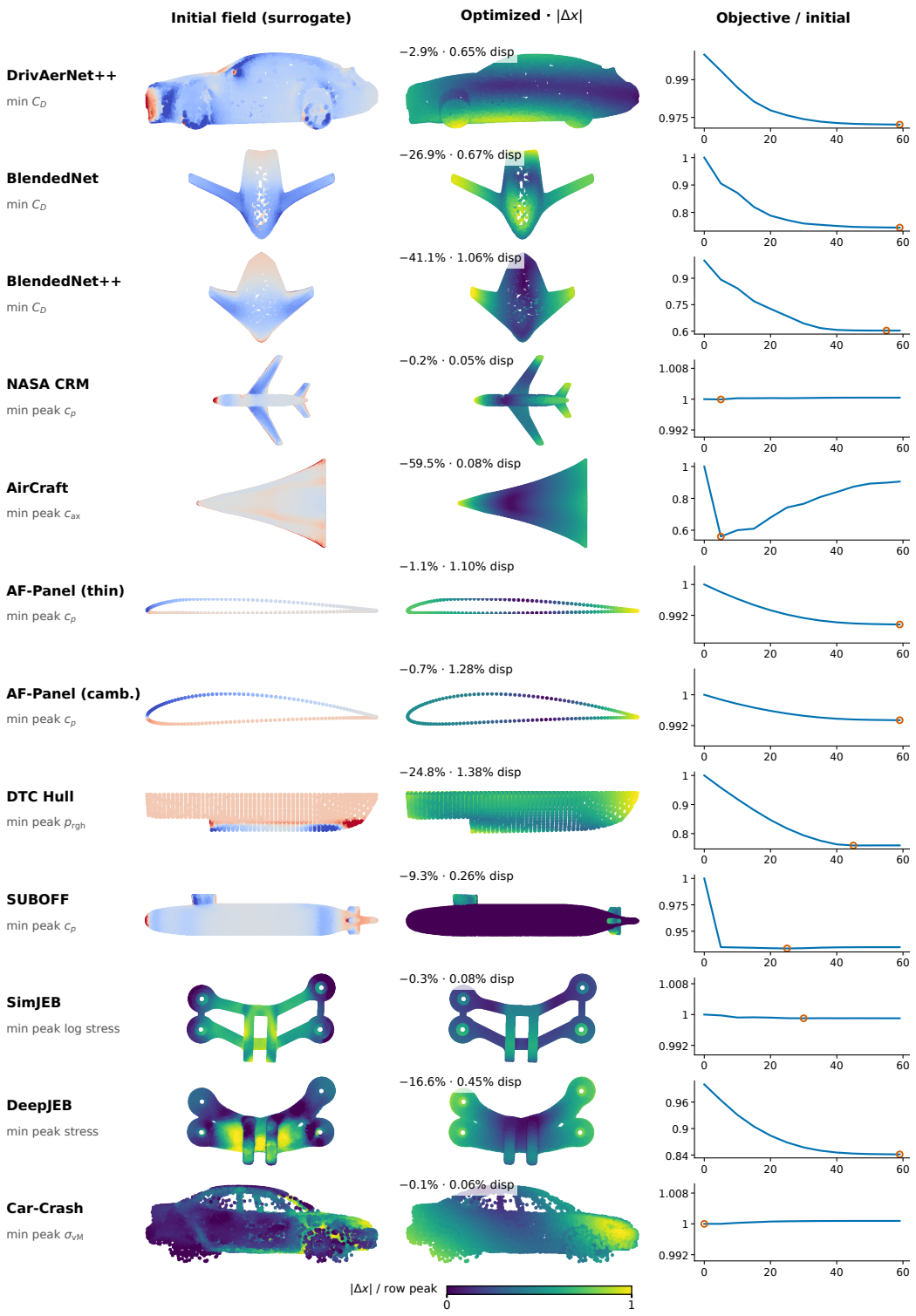


Figure 10: Inverse products of the twelve library compilations not shown in the main text, run under their registered objectives on the forward median cases. The columns show the initial surrogate field, the optimized geometry shaded by displacement over the row peak (badges: improvement and peak displacement), and the free-model objective at 5-step checkpoints, normalized to its initial value (open marker: minimum). Near-flat traces use a fixed $\pm 1\%$ span. Objectives insensitive to in-budget deformation report small gains (NASA CRM, SimJEB, Car-Crash).