SHIFT: A HIGH-FIDELITY BENCHMARK SUITE FOR 3D INDUSTRIAL FLUID DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

A majority of investigations involving fluid flow across science and engineering utilize Computational Fluid Dynamics (CFD) to simulate these flows. Traditional CFD is computationally expensive, driving interest in fast neural surrogates. However, existing benchmarks provide limited coverage of the combination of high-fidelity 3D geometries, complex flow physics and engineering quantities needed to systematically evaluate geometry-aware models in industrial settings. We introduce the SHIFT (Simulation for High Fidelity Training) benchmark suite, providing thousands of high-fidelity simulation results. Featuring both transient Delayed Detached Eddy Simulations (DDES) and steady Reynolds Averaged Navier Stokes (RANS) simulations for external aerodynamics, developed in partnership with domain experts and validated against experimental measurements for surface and flow quantities, SHIFT provides a standardized evaluation framework with predefined In-Distribution and Out-Of-Distribution sets to assess generalization to unseen geometries. Additionally, we report evaluations from several recent neural operator models, such as AB-UPT, DoMINO, Transolver, GeoTransolver; across engineering quantities of interest and provide baseline results that characterize the capabilities and limitations of these models.

1 INTRODUCTION

Computational Fluid Dynamics (CFD) is widely used in modern engineering, supporting design and optimization across applications ranging from automotive aerodynamics to aircraft design. However, traditional CFD is bottlenecked by substantial computational costs. Achieving high physical fidelity, such as resolving turbulent wake structures or capturing shock waves, requires iteratively solving complex, non-linear partial differential equations over massive 3D volumetric meshes with millions of degrees of freedom. For industrial-scale geometries, high-fidelity simulations can require substantial computational resources and may take days to weeks on High-Performance Computing systems (Caron et al., 2025; Slotnick et al., 2014; Probst et al., 2019). To overcome this limitation, geometry aware neural surrogate models have rapidly gained traction across the industry. By learning to approximate the underlying physics from data, such deep learning models bypass traditional iterative solvers, accelerating the time to prediction by orders of magnitude and potentially reducing prediction costs from expensive iterative simulations to near real-time inference.

In spite of the rapid evolution of deep learning architectures, with recent models now capable of processing unstructured CFD meshes with hundreds of millions of degrees of freedom, the rigorous evaluation of these surrogates is constrained by the available benchmark data. Existing benchmarks provide limited coverage of large-scale, high-fidelity 3D simulations with sufficient geometric diversity and engineering relevance. Researchers have to rely on simplified 2D flows, idealized "toy" geometries or limited fidelity simulation datasets for benchmarking. These datasets remain valuable for method development but often provide limited exposure to complex 3D flow phenomena encountered in engineering applications such as turbulent wake separation and transonic shock formation. Without benchmarks that capture realistic geometric and flow complexity, it can be difficult to determine whether a model generalizes to challenging engineering configurations or if it primarily benefits from the regularities of a simplified benchmark.

To address this need, we introduce the SHIFT (Simulation for High-Fidelity Training) dataset suite. The broader SHIFT datasets effort spans several scientific and engineering domains. Here, we fo-

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.

cus exclusively on 3D external aerodynamics through the SHIFT-SUV and SHIFT-Wing datasets. Together, these two datasets provide thousands of high-fidelity simulations of industrially relevant geometries, capturing diverse physical phenomena from automotive wake dynamics to steady state transonic shocks observed in flow fields around aerospace vehicles, designed to evaluate geometry-aware neural surrogate models under complex 3D flow conditions.

This paper establishes SHIFT as a benchmark suite for evaluating geometry-aware neural surrogate models on high-fidelity 3D external aerodynamics. To support reproducible benchmarking, we define deterministic In-Distribution training, validation and test splits. Additionally, we provide dedicated Out-Of-Distribution test sets that test generalization to unseen geometries. We construct a comparative benchmark spanning point-wise surface fields and volumetric flow fields. We evaluate several recent geometry-aware neural operator models (DoMINO (Ranade et al., 2025), Transolver (Wu et al., 2024), GeoTransolver (Adams et al., 2025) and AB-UPT (Alkin et al., 2025)) against this framework, we establish a reference baseline for the community. This resulting benchmark provides a common basis for identifying the regimes and quantities on which current 3D physics-AI models perform well or exhibit larger prediction errors.

2 Related Work

In the recent past, datasets like PDEBench (Takamoto et al., 2022), BlastNet (Chung et al., 2023), BubbleML (Hassan et al., 2023), AircraftVerse (Cobb et al., 2023), Eagle (Janny et al., 2023), etc have advanced the development of neural surrogates in specialized disciplines of Fluid Mechanics. Focusing on datasets for automobile aerodynamics, the AhmedML (Ashton et al., 2024b) dataset provides hybrid Reynolds-Averaged Navier-Stokes (RANS)/Large-Eddy Simulation (LES) data over 500 variations of the standard Ahmed Body car design. The WindsorML (Ashton et al., 2024a) dataset advances the realism of the automotive model, providing high fidelity LES data over 355 variations on the Windsor body used as a standard in Automotive CFD Prediction Workshops. The DrivAerML (Ashton et al., 2024c) dataset further increases geometric and simulation realism. Based on the DrivAer notchback geometry, it simulated 484 different variants using hybrid RANS-LES. The recent DrivAerNet++ (Elrefaie et al., 2024) dataset advanced this in terms of volume by covering over 8000 variants of fastback, notchback and estateback automotive designs using lower fidelity $k-\omega$ SST based simulations. These lower-fidelity simulations provide substantially larger coverage but may exhibit differences from higher-fidelity simulations in quantities such as flow separation, aerodynamic drag and wake structure (Ashton & Revell, 2015; Ashton et al., 2016).

In a similar vein, for external aerodynamics over aircraft, the AirfRANS (Bonnet et al., 2022) and the DeepCFD (Ribeiro et al., 2020) datasets, offer simulation data for thousands of airfoil designs, however, these focus on simplified 2D geometries. On the other hand, the Aerodynamic Design Optimization Dataset (Lee et al., 2015) focuses on high fidelity hybrid RANS/LES simulations of variations based on the NASA Common Research Model (CRM) design, but contains only a few hundred simulations.

Across these sub-fields, existing benchmarks illustrate a recurring tradeoff between dataset scale and simulation fidelity: large datasets often use lower-fidelity simulations whereas higher-fidelity datasets typically provide fewer configurations.

3 The SHIFT Benchmark Datasets

To train and evaluate deep learning architectures for fluid dynamics, datasets should complement simplified test cases with benchmarks that capture engineering complexity. The SHIFT benchmark suite was engineered specifically to address the scaling and evaluation needs of the machine learning community by adhering to five core design principles:

1. *Scale*: Recent geometry aware models (such as GeoTransolver with over 30 million trainable parameters) operate on high-dimensional meshes and thus, benefit from sufficiently large and diverse training corpora. To this end, SHIFT provides both numerous geometries and high-resolution meshes. For example, SHIFT-SUV contains roughly 4000 unique sample geometries, with each sample typically comprising approximately 45 million cells for volume data and

millions of points for surface representation. At each volume cell and surface point, multiple floating-point values representing physical quantities such as velocity and pressure are provided.

2. *Engineering Relevance*: Benchmarks must advance past simplified 2D simulations, toy geometries like flow over periodic hills or overly simplified boundary conditions. SHIFT exposes models to simulation setups that incorporate engineering-relevant boundary conditions and geometric details. For instance, in SHIFT-SUV, engineering-relevant physics such as a rolling floor and wheel rotation are explicitly modeled via translating and rotating surface boundary conditions. In this spirit, both these SHIFT datasets were developed in partnership with domain experts.
3. *Physical Fidelity*: In the machine learning community, "fidelity" is often associated with spatial resolution or grid density. In computational fluid dynamics, fidelity is broader: it refers to how accurately relevant physics are represented, and can be influenced through both spatial resolution and turbulence modeling. The appropriate route to higher fidelity depends on the flow regime. SHIFT-Wing, for example, uses solution-adaptive RANS to concentrate resolution around thin features such as shock waves and boundary layers. By contrast, flows dominated by unsteady separation and strongly anisotropic wakes require greater fidelity in the turbulence treatment itself. Most existing ML-CFD datasets rely on two-equation RANS models such as $k-\omega$ SST. While these models can accurately capture many flow features when sufficiently resolved, they model all turbulent scales and rely on simplifying assumptions such as the linear eddy-viscosity hypothesis. For complex bluff-body aerodynamics, this can lead to errors in separation and reattachment (Spalart, 2000), wake structure (Wilcox et al., 1998), and integrated quantities such as drag (Rüttgers et al., 2019; Ashton et al., 2016). SHIFT-SUV therefore uses transient DDES, resolving large, energy-containing turbulent structures while modeling smaller scales. Its time-averaged reference solutions provide richer representations of separated and unsteady wake physics for training and evaluating neural surrogates.
4. *Granularity and Modality*: ML models must be capable of predicting both highly localized fields and integrated engineering quantities. SHIFT provides both surface and volumetric fields (e.g., pressure, velocity, shear stress), enabling the evaluation of both point-wise predictions and integrated Quantities of Interest (QoIs) like drag and lift coefficients.
5. *Geometric Diversity*: Comprehensive testing requires evaluating intra-dataset parametric variance and inter-dataset topological shifts. SHIFT achieves this through a vast range of parameterizations. For instance, SHIFT-SUV covers four distinct configurations (full-scale estate, full-scale fastback, quarter-scale estate, quarter-scale fastback). We include smooth and detailed underbodies in each of these classes along with a wide range of parameterizations within each class. Robust evaluation requires testing both variation within the training geometry family and shifts to geometries outside that family. SHIFT-SUV provides within-family variation across vehicle configurations and also includes different geometries as an Out-Of-Distribution evaluation set.

Finally, SHIFT defines standardized evaluation protocols, including In-Distribution splits and dedicated Out-Of-Distribution sets, together with reference evaluations of recent physics-AI models.

3.1 SHIFT-SUV: Automotive Aerodynamics

SHIFT-SUV contains nearly 4000 transient Delayed Detached Eddy Simulation (DDES) cases generated from geometric variants of the AeroSUV platform (Zhang et al., 2019), an open-source SUV reference geometry developed by FKFS[1] that shares lineage with the DrivAer notchback (Heft et al., 2012) and extends it into the SUV form factor. The dataset spans four configurations: full- and quarter-scale models, each in estate and fastback variants (Figure 1). Two scales of model are included due to relevance to automotive manufacturers, who test designs at quarter-scale before manufacturing full-sized models. The 3958 full-scale samples form the training corpus. The additional 1997 quarter-scale samples are reserved for validating models against FKFS wind-tunnel measurements (Zhang, 2021), which provided delta in forces between configurations, shown in Table 1.

Additional validation is based on results at the AutoCFD4 workshop (Economon et al., 2024; Hupertz et al., 2022), an automotive industry benchmarking exercise for CFD solvers. Extensive full-scale vehicle experimental data comparisons have been performed on the DrivAer, an automotive sedan model from which the AeroSUV is derived (Zhang et al., 2019). Extensive validation on this

[1]Forschungsinstitut für Kraftfahrwesen und Fahrzeugmotoren, Stuttgart

AeroSUV Drag (C_D)	Medium Mesh (48M)	Fine Mesh (152M)	Experiment
Estate configuration	0.331	0.331	-
Fastback configuration	0.289	0.288	-
Δ between configurations	0.042	0.044	0.027

Table 1: Quarter-scale AeroSUV drag coefficients demonstrating mesh insensitivity of the quantities of interest.

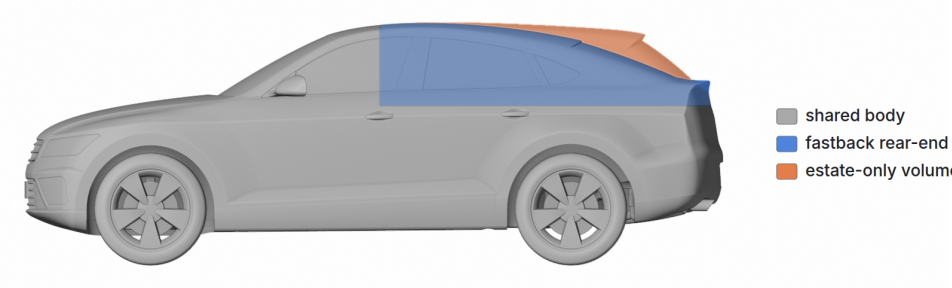


Figure 1: The two reference AeroSUV configurations from which SHIFT-SUV is generated and which validation was performed. Each is sampled at full and quarter scale, yielding four configurations.

vehicle is provided in the references, with reproductions of comparisons of integrated forces and point-wise profiles with experimental data included in the manuscript in Section C.5.

Engineering-relevant boundary conditions are modeled explicitly: a translating-floor and rotating-wheel surfaces represent road-vehicle interaction, with a uniform $31.29\ \mathrm{m/s}$ freestream representing highway cruise. Details of the simulation software, boundary conditions, and workflow are provided in Section C.1.

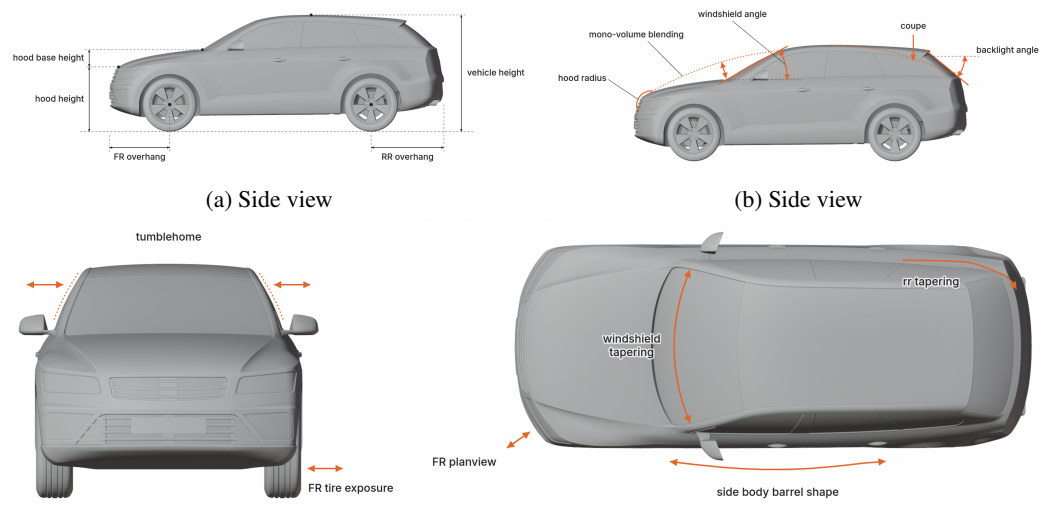
Geometric variants of the reference configurations are produced by non-orthogonal morphing boxes in the ANSA software, details shown in Figure A.2, parameterized by sixteen morphing variables informed in partnership with Honda Motors. The parameters describe shape changes familiar to vehicle designers (hood height, windshield angle, ride height, rear tapering and others) and intentionally share control points so that multiple parameters jointly modify overlapping regions. Geometry common to all variants (wheels, tires, suspension components) are excluded from the morph and held invariant across the dataset. A visual diagram of the parameter directions is shown in Figure 2, however the full sixteen-parameter table with value ranges is deferred to Table A.1.

The released benchmark additionally includes ten truck geometries with their corresponding simulations at the same operating conditions as the SUV results. These are excluded from training and validation and are used exclusively for OOD evaluation for geometry distribution shift.

SHIFT-SUV provides both integrated engineering quantities and high-dimensional spatial fields. For each case, the simulation mesh and geometry representation, together with the corresponding time-averaged surface fields (pressure and wall shear-stress), volumetric flow fields (pressure and the velocity vector) and integrated aerodynamic quantities are provided. The fields are time-averaged over 0.275 s and 2.75 s for quarter and full scale, respectively.

3.2 SHIFT-Wing: Transonic Aircraft Aerodynamics

SHIFT-Wing populates a parametric design space around a high-speed wing-body configuration reconstructed from the NASA Common Research Model (CRM) (NASA Common Research Model Project), a reference platform extensively studied through the AIAA Drag Prediction Workshop series (Tinoco et al., 2018). We construct a parametric model in OnShape that exposes seven macroscopic "designer-friendly" variables (Table 3): wing aspect ratio, $^1/_4$-chord sweep angle, root-chord extension factor, fuselage diameter, and three twist parameters spanning the root, the root-to-Yehudi-break segment, and the break-to-tip segment. Two flow parameters are co-sampled with geometry:



(a) Side view (b) Side view

(c) front view (d) top view

Figure 2: Visualizations of the SHIFT-SUV morphing parameters described in Table A.1.

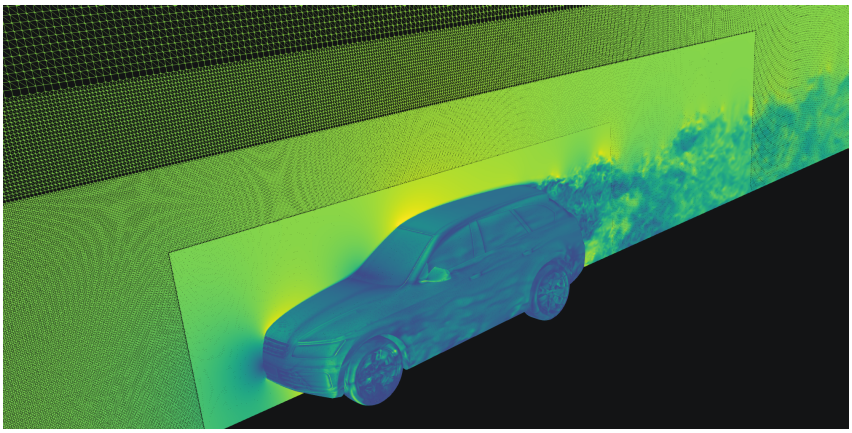


Figure 3: Representative SHIFT-SUV solution field: instantaneous velocity contours on the centerplane and wall shear stress on the vehicle surface for a full-scale Estate variant. The transient DDES resolves the unsteady wake structure downstream of the vehicle. The computational mesh density is depicted by the centerplane wireframe.

angle of attack and Mach number. Together these parameters cover sub-critical through transonic regimes (Figures 4 and 5).

An Out-Of-Distribution set is included consisting of 10 geometries with the shape parameters extrapolated to 1–5% of their range outside the In-Distribution design space, each simulated at $\mathrm{Ma} = 0.5$ and 0.85 at $4.0°$ angle of attack, for a total of 20 cases. All parameters and ranges were determined in consultation with domain experts from Otto Aerospace.

The dataset is generated in batches at fixed Mach number, with geometry parameters and angle of attack sampled by Latin Hypercube within each batch. The Mach number strongly influences the presence and strength of compressibility effects and shock structures, so fixed-Mach batches provide a controlled way to evaluate models across distinct flow regimes. Surrogate models can therefore be evaluated by regime, for example, subsonic and transonic cases, rather than only on an aggregated test set. To resolve sharp shock features without manually pre-meshing each variant, the dataset uses solution-adaptive meshing (Krakos et al., 2025), which iteratively refines the mesh anisotropically according to solution gradients. The resulting meshes vary substantially in node count, connectivity

and spatial resolution between samples. These adapted meshes define the high-resolution CFD target discretization and therefore influence the spatial distribution of training targets, with greater sampling density in regions containing features such as shocks and shear layers. Importantly, the adapted mesh is not required as an input at inference: the released benchmark also provides a CAD-direct, approximately isotropic surface tessellation and solution-independent query representations for evaluating geometry-to-field prediction.

As this mesh adaptation is driven by the evolving CFD solution, the adapted discretization can contain information about where flow features occur. We distinguish between information used to construct the training targets and information available to the surrogate at inference. SHIFT-Wing does not require the solution-adapted mesh at inference. The benchmark includes both the adapted CFD discretization and a CAD-direct isotropic tessellation, representative node distributions are shown in Figure A.8 for the same wing, enabling this distinction to be evaluated empirically. This distinction has also been evaluated independently on SHIFT-Wing. Alkin et al. (2025) trained AB-UPT using solution-adapted target meshes while supplying only an isotropic CAD representation at inference. Their CAD-input variant achieved $R^2 = 1.00$ for both drag and lift, with median errors of 0.2% and 0.3%, respectively, and maximum errors of 3.1% and 5.0%. They further report approximately 0.6 s inference time for drag and lift from a 200k-point isotropic CAD representation. These results provide evidence that accurate inference from a solution-independent geometry representation is possible despite training on solution-adapted targets. To enable controlled studies of discretization dependence, each SHIFT-Wing case includes the final solution-adapted mesh, a CAD-direct approximately isotropic surface tessellation and intermediate meshes from the adaptation sequence. This provides a mesh-density and discretization ladder with which users can quantify sensitivity to solution-dependent spatial sampling.

This solution approach has been independently exercised at the 5th AIAA High Lift Prediction Workshop (HLPW) (Krakos et al., 2025): verification on the CRM-HL Wing-Body configuration (a CRM geometry related to the one underpinning SHIFT-Wing) and validation against experimental data on increasingly complex configurations, providing an external, regime-distinct validation of the solver technology applied to SHIFT-Wing. Grid convergence of this approach is shown in Table 2, with additional validation results deferred to Section C.9.

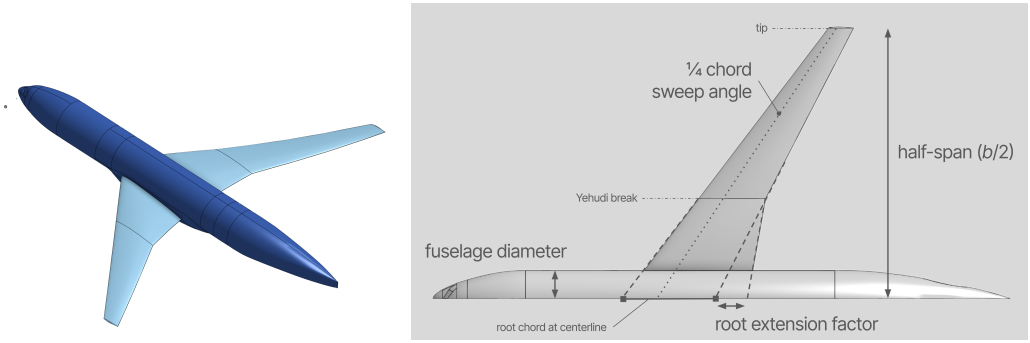
mesh size n	C_L	C_D	C_M
4,573,022	0.99949	0.0765120	-0.000125
9,339,419	1.07353	0.0649825	-0.063409
18,641,291	1.07686	0.0642010	-0.065502
35,634,069	1.07788	0.0639613	-0.066137
72,132,382	1.07793	0.0638114	-0.066248
145,236,498	1.07832	0.063768	-0.066366

Table 2: Adaptive grid convergence results for HLPW Case 1 (CRM-HL Wing-Body configuration), using the standard SA turbulence model.

This dataset provides a challenging setting for neural surrogates because the reference discretization varies substantially across samples in node count, connectivity and spatial resolution. The solution-adaptive procedure concentrates resolution around flow features such as shocks and stagnation regions, while the released solution-independent geometry representations enable models to be evaluated without requiring a precursor CFD solution at inference. The resulting benchmark therefore supports controlled studies of both field prediction and sensitivity to discretization.

3.3 Data Generation Details

All CFD reference solutions in SHIFT are produced with a GPU-native finite-volume code that is second-order in both space and time (Economon et al., 2024; Krakos et al., 2025). Using the same solver framework across both datasets reduces implementation-level confounders, including differences in numerical discretization, time integration and GPU execution when comparing surrogate-model performance across the automotive and aerospace benchmarks. The two datasets exercise the solver under deliberately different regimes. SHIFT-SUV uses approximately 45M-cell hex-dominant meshes generated per-sample in ANSA, with transient DDES on a Spalart–



(a) Assembled wing-body geometry

(b) Macro-parameter diagram

Figure 4: SHIFT-Wing geometry: full-span wing-body configuration constructed from the parametric CRM model (a) & the 7 design parameters defining the planform sampling space (b).

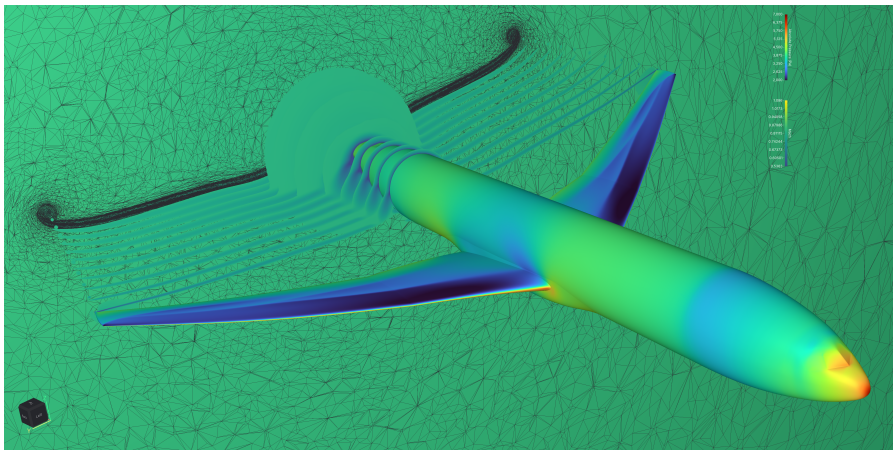
Design space parameter	Symbol	Min	Reference	Max
Aspect ratio	AR	7.5	9.0	11.0
$^1/_4$-chord sweep angle	$\beta_{\frac{1}{4}}$	$25°$	$35°$	$37.5°$
Root-chord extension factor	F_{C_r}	1.0	1.373	1.4
Fuselage diameter	D_f	240 in	240 in	258 in
Root twist	t_r	$3°$	$6.717°$	$9°$
Δ twist root-to-break	$\Delta t_{r,b}$	$-7°$	$-5.953°$	$-3°$
Δ twist break-to-tip	$\Delta t_{b,t}$	$-7.5°$	$-4.513°$	$-1.5°$
Angle of attack	α	$0°$	—	$4°$
Mach number	Ma	0.5	—	0.85

Table 3: SHIFT-Wing design space: Rows above the rule define the parametric geometry; the two rows below define the flow conditions. Mach number is sampled in fixed batches; all other variables use Latin Hypercube sampling within each Mach batch.

Allmaras backbone (Spalart et al., 2006); a hybrid centered/upwind convective scheme with proprietary blending suppresses dissipation in resolved-LES regions, and an advanced shielding function prevents modeled-stress depletion that would otherwise produce artificially early flow separation. SHIFT-Wing uses steady RANS with Spalart–Allmaras combined with automatic mesh adaptation, an anisotropic solution-adaptive technology that refines the mesh along emerging flow features without user intervention. The result is a benchmark suite that spans two distinct learning challenges: in SHIFT-SUV the released reference fields originate from fixed-resolution input discretizations and the model must learn wake dynamics, while in SHIFT-Wing the reference fields are defined on solution-adaptive, variable-topology meshes, and the benchmark also provides solution-independent geometry representations for inference. Detailed turbulence-modeling, meshing and validation specifics are provided in Section C.

4 Evaluation & Analysis

We define fixed training, validation and test splits for both datasets to provide a consistent evaluation protocol. For SHIFT-SUV, the benchmark uses 3,958 full-scale simulations spanning the Estate and Fastback configurations, partitioned into 80/10/10 training, validation and test sets. For SHIFT-Wing, the benchmark uses 3,941 simulations spanning subsonic (Mach 0.5; 1,975 simulations) and transonic (Mach 0.85; 1,966 simulations) regimes, using the same 80/10/10 partition. The splits are deterministic and fixed across baseline experiments to ensure reproducibility and consistent comparison. In addition to these In-Distribution splits, Out-Of-Distribution test sets evaluate generalization beyond the training distribution. By fixing these hold-out sets, we prevent data leakage and distinguish interpolation within the sampled design space from generalization to deliberately shifted geometries. To ensure strict benchmark integrity and reproducibility, we have independently executed & verified in-house runs for all models on the standardized data splits.



(a) Representative SHIFT-Wing solution at $\mathrm{Ma} = 0.85$: surface pressure contours on the airplane and a centerline cut of the solution-adapted mesh

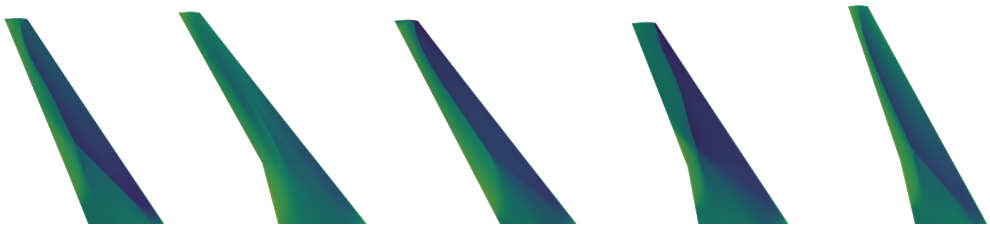


Figure 5: Top row: full solution visualization. Bottom row: five different realizations of the adapted SHIFT-Wing geometry and solution, visualizing the surface pressure field on the top surface of the wing.

We evaluate surrogate accuracy at both the field and engineering levels. For spatial fields, we report Relative L_1 error on surface pressure (p_s), wall shear stress (τ_w), volumetric pressure (p_v), and velocity (u). These quantities probe complementary aspects of the solution, including pressure distributions, boundary layer behavior and three-dimensional flow structure. As wall shear stress approaches zero in regions of weak surface loading, its Relative L_1 error should be interpreted with care. We treat it primarily as a field-reconstruction metric rather than as a measure of aerodynamic performance. We benchmark four recent neural-operator architectures representing different approaches to geometry-aware 3D field prediction: DoMINO (Ranade et al., 2025), Transolver (Wu et al., 2024), GeoTransolver (Adams et al., 2025) and AB-UPT (Alkin et al., 2025). DoMINO uses multi-scale local geometric queries, Transolver uses physics-aware attention, GeoTransolver combines physics-aware attention with multi-scale geometry-aware embeddings and AB-UPT uses anchored geometry tokens with a branched architecture. These models provide a representative baseline set for evaluating field prediction on irregular 3D aerodynamic domains.

Table 4: Baseline field evaluation (relative L_1 errors in $\%$) across the SHIFT-SUV Estate and Fastback testing splits.

	Estate				Fastback			
	Surface		Volume		Surface		Volume	
Model	p_s	τ_w	p_v	$\mathbf{u}$	p_s	τ_w	p_v	$\mathbf{u}$
DoMINO	0.0100	12.24	0.0062	8.14	0.0100	11.74	0.0067	7.73
Transolver	0.0079	4.98	0.0040	1.87	0.0078	4.97	0.0039	1.81
AB-UPT	0.0064	4.95	0.0025	2.25	0.0064	5.03	0.0024	2.21
GeoTransolver	0.0057	3.81	0.0026	1.36	0.0056	3.70	0.0023	1.30

Table 4 reports Relative L_1 field errors for the SHIFT-SUV Estate and Fastback test splits. The four models show broadly similar trends across the two geometries, with velocity and volumetric pressure generally predicted more accurately than wall shear stress. GeoTransolver obtains the lowest reported velocity error on both configurations (1.36% on Estate and 1.30% on Fastback), while AB-UPT has a slightly lower volumetric-pressure error on Estate (0.0025% versus 0.0026%). Wall shear stress remains substantially more difficult than the other fields across all models, with errors of approximately 4–12%. These results indicate that boundary-layer quantities remain challenging even when bulk flow fields are predicted accurately. The relatively consistent trends between Estate and Fastback also provide a useful check on model behavior across the two vehicle configurations.

Table 5 reports Relative L_1 field errors for SHIFT-Wing at Mach 0.5 and Mach 0.85. Surface-pressure errors generally increase at Mach 0.85, particularly for DoMINO, consistent with the greater difficulty of representing sharp transonic shock features. In contrast, the velocity errors of GeoTransolver remain similar across the two regimes (1.92% at Mach 0.5 and 2.00% at Mach 0.85), while Transolver shows a similarly small change (3.60% to 3.72%). AB-UPT obtains the lowest surface-pressure error at Mach 0.85 (0.079%), compared with 0.081% for GeoTransolver, while GeoTransolver obtains the lowest volumetric-pressure error (0.099%). Wall shear stress remains challenging for all four models, with errors of approximately 10–13% across both regimes. DoMINO has substantially larger volumetric-velocity errors than the other models, particularly at Mach 0.85, while the transformer-based models show smaller changes across the two Mach conditions. Overall, the results indicate that transonic shock structure and wall-bounded quantities remain challenging targets even when volumetric flow fields are predicted with relatively low error.

Table 5: Baseline field evaluation (relative L_1 errors in $\%$) across the SHIFT-Wing testing splits at Mach Numbers 0.5 (subsonic flow regime) and 0.85 (transonic flow regime).

	Ma $= 0.5$				Ma $= 0.85$			
	Surface		Volume		Surface		Volume	
Model	p_s	τ_w	p_v	$\mathbf{u}$	p_s	τ_w	p_v	$\mathbf{u}$
DoMINO	0.468	10.20	2.250	21.34	1.880	13.35	3.170	29.20
Transolver	0.094	12.40	1.190	3.60	0.098	13.20	1.340	3.72
AB-UPT	0.022	12.50	0.027	9.56	0.079	13.30	0.125	9.51
GeoTransolver	0.021	12.20	0.022	1.92	0.081	13.01	0.099	2.00

5 Conclusions

In this work, we introduced SHIFT, a benchmark suite for evaluating geometry-aware neural surrogates on high-fidelity 3D industrial external aerodynamics. SHIFT-SUV provides thousands of DDES simulations of parametrically varied vehicle geometries, while SHIFT-Wing provides thousands of steady transonic RANS simulations with solution-adaptive discretizations. Together, the datasets test surrogate models on geometric variation, complex flow physics, high-dimensional spatial fields and substantial variation in CFD discretization. We define standardized training, validation and test protocols, together with dedicated Out-Of-Distribution evaluation settings for geometry shifts. The benchmark includes surface and volumetric fields as well as engineering quantities of interest, enabling evaluation at both field and engineering levels. Baseline results show that current neural operators can predict bulk flow fields with relatively low error while wall shear stress and sharp transonic features remain substantially more difficult. These results establish reference points for future methods and provide a controlled testbed for studying geometry generalization, discretization dependence and engineering-level accuracy in 3D CFD surrogates.

AI use statement

We used large language model (LLM) tools to obtain feedback on the manuscript, primarily to identify opportunities for improving clarity and presentation. The feedback was used in accordance with the ICLR author guidelines to improve the manuscript. LLMs were not used to develop the scientific methodology, formulate hypotheses, design or conduct experiments, analyze or interpret

scientific results, generate scientific claims or produce the technical contributions presented in this work. The authors independently conducted and verified all scientific work and are fully responsible for the content, claims and conclusions of this paper.

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.

Benedikt Alkin, Maurits Bleeker, Richard Kurle, Tobias Kronlachner, Reinhard Sonnleitner, Matthias Dorfer, and Johannes Brandstetter. AB-UPT: Scaling neural CFD surrogates for high-fidelity automotive aerodynamics simulations via anchored-branched universal physics transformers. *arXiv preprint arXiv:2502.09692*, 2025.

Neil Ashton and Alistair Revell. Key factors in the use of ddes for the flow around a simplified car. *International Journal of Heat and Fluid Flow*, 54:236–249, 2015.

Neil Ashton, Alastair West, Sylvain Lardeau, and Alistair Revell. Assessment of rans and des methods for realistic automotive models. *Computers & fluids*, 128:1–15, 2016.

Neil Ashton, Jordan B Angel, Aditya S Ghate, Gaetan K Kenway, Man L Wong, Cetin Kiris, Astrid Walle, Danielle C Maddix, and Gary Page. Windsorml: High-fidelity computational fluid dynamics dataset for automotive aerodynamics. *Advances in Neural Information Processing Systems*, 37:37823–37835, 2024a.

Neil Ashton, Danielle C Maddix, Samuel Gundry, and Parisa M Shabestari. Ahmedml: High-fidelity computational fluid dynamics dataset for incompressible, low-speed bluff body aerodynamics. *arXiv preprint arXiv:2407.20801*, 2024b.

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

Philipp Bekemeyer, Nathan Hariharan, Andrew M Wissink, and Jason Cornelius. Introduction of applied aerodynamics surrogate modeling benchmark cases. In *AIAA SCITECH 2025 Forum*, pp. 0036, 2025.

Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. Airfrans: High fidelity computational fluid dynamics dataset for approximating reynolds-averaged navier–stokes solutions. *Advances in Neural Information Processing Systems*, 35:23463–23478, 2022.

Timothy R Brooks, Gaetan K Kenway, and Joaquim RRA Martins. Undeflected common research model (ucrm): An aerostructural model for the study of high aspect ratio transport aircraft wings. In *35th AIAA Applied Aerodynamics Conference*, pp. 4456, 2017.

Timothy R Brooks, Gaetan KW Kenway, and JRRA Martins. Ucrm: an aerostructural model for the study of flexible transonic aircraft wings. *AIAA Journal*, 2018.

Clement Caron, Philippe Lauret, and Alain Bastide. Machine learning to speed up computational fluid dynamics engineering simulations for built environments: A review. *Building and Environment*, 267:112229, 2025.

Wai Tong Chung, Bassem Akoush, Pushan Sharma, Alex Tamkin, Ki Sung Jung, Jacqueline Chen, Jack Guo, Davy Brouzet, Mohsen Talei, Bruno Savard, et al. Turbulence in focus: Benchmarking scaling behavior of 3d volumetric super-resolution with blastnet 2.0 data. *Advances in Neural Information Processing Systems*, 36:77430–77484, 2023.

Adam Cobb, Anirban Roy, Daniel Elenius, Frederick Heim, Brian Swenson, Sydney Whittington, James Walker, Theodore Bapty, Joseph Hite, Karthik Ramani, et al. Aircraftverse: A large-scale multimodal dataset of aerial vehicle designs. *Advances in Neural Information Processing Systems*, 36:44524–44543, 2023.

Thomas Economon, Darren Engwirda, Jonathan Ho, Michael Mara, and Raghav Mansharamani. Luminary contribution to the autocfd4 workshop, 2024. URL https://autocfd.org/autocfd4/.

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. *Advances in Neural Information Processing Systems*, 37:499–536, 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Sheikh Md Shakeel Hassan, Arthur Feeney, Akash Dhruv, Jihoon Kim, Youngjoon Suh, Jaiyoung Ryu, Yoonjin Won, and Aparna Chandramowlishwaran. Bubbleml: A multiphase multiphysics dataset and benchmarks for machine learning. *Advances in Neural Information Processing Systems*, 36:418–449, 2023.

Angelina I Heft, Thomas Indinger, and Nikolaus A Adams. Introduction of a new realistic generic car model for aerodynamic investigations. Technical report, SAE Technical Paper, 2012.

Burkhard Hupertz, Neil Lewington, Charles Mockett, Neil Ashton, and Lian Duan. Towards a standardized assessment of automotive aerodynamic CFD prediction capability-AutoCFD 2: Ford drivaer test case summary. Technical report, SAE Technical Paper, 2022.

Steeven Janny, Aurélien Beneteau, Madiha Nadri, Julie Digne, Nicolas Thome, and Christian Wolf. Eagle: Large-scale learning of turbulent fluid dynamics with mesh transformers. *arXiv preprint arXiv:2302.10803*, 2023.

Joshua Krakos, Pedro Gomes, Jonathan Ho, Michael Mara, Gonzalo Sáez, Adrien Loseille, Thomas Economon, and Juan J Alonso. Gpu-based and adaptive solution technology for the 5th aiaa high lift prediction workshop. In *AIAA SCITECH 2025 Forum*, pp. 0496, 2025.

Christopher Lee, David Koo, Karla Telidetzki, Howard Buckley, Hugo Gagnon, and David W Zingg. Aerodynamic shape optimization of benchmark problems using jetstream. In *53rd AIAA Aerospace Sciences Meeting*, pp. 0262, 2015.

NASA Common Research Model Project. High speed common research model (crm). https://commonresearchmodel.larc.nasa.gov/home-2/high-speed-crm/.

George Panagos and Nikolas Drivakos. Case study on the FE-model and geometry morphing in structural analysis. https://www.beta-cae.com/events/c2pdf/H2-3-1.pdf.

Axel Probst, Tobias Knopp, Cornelia Grabe, and Jens Jägersküpper. Hpc requirements of high-fidelity flow simulations for aerodynamic applications. In *European Conference on Parallel Processing*, pp. 375–387. Springer, 2019.

Rishikesh Ranade, Mohammad Amin Nabian, Kaustubh Tangsali, Alexey Kamenev, Oliver Hennigh, Ram Cherukuri, and Sanjay Choudhry. Domino: A decomposable multi-scale iterative neural operator for modeling large scale engineering simulations. *arXiv preprint arXiv:2501.13350*, 2025.

Mateus Dias Ribeiro, Abdul Rehman, Sheraz Ahmed, and Andreas Dengel. Deepcfd: Efficient steady-state laminar flow approximation with deep convolutional neural networks. *arXiv preprint arXiv:2004.08826*, 2020.

Mario Rüttgers, Junshin Park, and Donghyun You. Large-eddy simulation of turbulent flow over the drivaer fastback vehicle model. *Journal of Wind Engineering and Industrial Aerodynamics*, 186: 123–138, 2019.

Jeffrey P Slotnick, Abdollah Khodadoust, Juan Alonso, David Darmofal, William Gropp, Elizabeth Lurie, and Dimitri J Mavriplis. CFD vision 2030 study: a path to revolutionary computational aerosciences. Technical report, 2014.

Philippe R Spalart. Strategies for turbulence modelling and simulations. *International journal of heat and fluid flow*, 21(3):252–263, 2000.

Philippe R Spalart, Shur Deck, Michael L Shur, Kyle D Squires, M Kh Strelets, and Andrei Travin. A new version of detached-eddy simulation, resistant to ambiguous grid densities. *Theoretical and computational fluid dynamics*, 2006.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: An extensive benchmark for scientific machine learning. *Advances in neural information processing systems*, 35:1596–1611, 2022.

Edward N Tinoco, Olaf P Brodersen, Stefan Keye, Kelly R Laflin, Edward Feltrop, John C Vassberg, Mori Mani, Ben Rider, Richard A Wahls, Joseph H Morrison, et al. Summary data from the sixth aiaa cfd drag prediction workshop: Crm cases. *Journal of Aircraft*, 55(4):1352–1379, 2018.

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*, pp. 6919, 2008.

David C Wilcox et al. *Turbulence modeling for CFD*, volume 2. DCW industries La Canada, CA, 1998.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries. In *ICML*, 2024.

Chenyi Zhang. *Aerodynamic study on the vehicle shape parameters with respect to ground simulation*. Springer, 2021.

Chenyi Zhang, Max Tanneberger, Timo Kuthada, Felix Wittmeier, Jochen Wiedemann, and Juliane Nies. Introduction of the aerosuv-a new generic suv model for aerodynamic research. Technical report, SAE Technical Paper, 2019.

APPENDIX

A DATA AVAILABILITY, HOSTING AND MAINTENANCE

Hosting and license. Both SHIFT-SUV and SHIFT-Wing are shared under the CC-BY-NC 4.0 license. While the machine learning community traditionally favors frictionless data distribution, the physical reasoning capabilities enabled by large-scale 3D CFD datasets present inherent dual use realities. Furthermore, the broader SHIFT dataset ecosystem explicitly includes datasets (for instance, SHIFT-CCA (UAV design) and SHIFT-Missile (high speed defense payloads)) that possess clear defense and military applications. Because the neural architectures evaluated in this benchmark are domain-agnostic, we utilize a credentialed access model that requires contact sharing and approval for the public release of the SHIFT datasets to ensure responsible data stewardship.

Furthermore, this credentialed access model provides vital telemetry on dataset utilization, directly informing our long-term resource allocation strategy. By verifying legitimate scientific inquiry and tracking usage trends, we can accurately identify which physical domains provide the most value to the ML-CFD community, allowing us to prioritize supercomputing resources toward expanding the most highly utilized datasets in future releases.

Reviewer sample and exploration harness. A representative subset of both datasets (one full case per main subset, about 5 GB) is publicly downloadable without credentials. The data loaders, preprocessing scripts, canonical split manifests, and an exploration notebook are released on GitHub at https://github.com/shift-benchmark-submission/shift-benchmark-submission, which also provides full instructions for accessing the datasets, documentation for all scripts, and copies of the Croissant metadata with RAI fields. The full datasets (approximately 29 TB) are hosted in an access-controlled bucket during review; reviewers receive a read-only access key in an OpenReview comment visible only to reviewers and area chairs, posted as soon as OpenReview permits and no later than the opening of the discussion period, and the repository's download script uses it directly.

Maintenance plan. We commit to the long-term maintenance of both datasets. New samples are added in semantically versioned releases, with prior versions retained for the reproducibility of published results. Bug reports, documentation issues, and metadata corrections are tracked publicly in the GitHub repository linked above. Substantive changes (samples added, fields renamed, license amendments) are recorded in a CHANGELOG and announced via the HuggingFace dataset cards as well as email notifications to registered users.

B SUMMARY STATISTICS FOR THE DATASETS

Table A.1 describes in detail the SHIFT-SUV input parameter distribution. SHIFT-SUV fixes the freestream dataset-wide ($U_\infty = 31.29$ m/s, $\rho_\infty = 1.204$ kg/m^3, $p_\infty = 101.3$ kPa, $T_\infty = 293.1$ K) with $A_{\text{ref}} = 2.464$ m^2 and $L_{\text{ref}} = 2.786$ m, so geometry is its only varying input.

The SHIFT-SUV design of experiments unlocks shape parameters in stages indexed by run number, so a parameter may be recorded at its default of zero because it was not yet active. Table A.1 therefore reports statistics over active cases only, and gives the active count for each parameter.

Tables A.2–A.4 report summary statistics for target quantities.

Per-case values, every split, and the complete column set are released as CSV in the dataset and described in the accompanying Croissant metadata.

C EXTENDED CFD METHODOLOGY

C.1 NUMERICAL SOLVER AND GPU IMPLEMENTATION

The flow solver used to generate ground-truth solutions for both SHIFT-SUV and SHIFT-Wing is a GPU-native, cell-centered finite-volume code operating on arbitrary polyhedral unstructured

Table A.1: SHIFT-SUV full-scale-4k: sampled design variables. Statistics are over cases in which the parameter is active.

Parameter	Active cases	Min	Max	Mean	Std
hood height (mm)	3958	-120.0	200.0	32.5	81.1
hood radius (mm)	3958	-160.0	160.0	0.0	81.8
rear tapering (mm)	3958	-200.0	400.0	89.9	152.4
windshield angle (mm)	3958	-480.0	480.0	-0.9	231.4
front planview (mm)	3762	-320.0	320.0	0.0	160.7
front tire exposure (mm)	3762	-440.0	240.0	-99.8	191.7
hood base height (mm)	3762	-80.0	200.0	48.5	73.6
vehicle height (mm)	3762	-200.0	200.0	0.0	101.3
mono-volume blending (mm)	3063	0.0	160.0	79.3	45.9
backlight angle (mm)	2966	-400.0	480.0	46.5	239.9
front overhang (mm)	1976	-200.0	200.0	0.5	115.4
rear overhang (mm)	1976	-400.0	400.0	-1.3	230.9
side-body barrel shape (mm)	1976	-120.0	120.0	-0.2	69.2
tumblehome (mm)	1976	-200.0	200.0	0.4	115.5
windshield planview (mm)	1976	-200.0	120.0	-39.6	92.4
coupe, estate only (mm)	988	-120.0	120.0	-0.4	69.4

Table A.2: SHIFT-SUV full-scale-4k (3958 cases): target quantities. Mean and Std pooled over all values. Low/High are median per-case p01/p99 for fields (to suppress single-cell outliers), min/max across cases otherwise.

Quantity	Low	High	Mean	Std
C_D	0.2462	0.5245	0.323	0.03582
C_L	-0.4262	0.2186	-0.09566	0.09424
Surface Pressure (Pa)	1.006e+05	1.019e+05	1.012e+05	269.5
Wall Shear Stress Magnitude (N/m^2)	0.01782	5.373	1.27	1.201
Surface cells	2.33M	2.94M	2.53M	6.928e+04
Surface points	2.34M	2.95M	2.55M	6.976e+04
Volume cells	44.03M	49.74M	46.17M	6.273e+05
Volume points	48.10M	54.41M	50.53M	7.059e+05
Volume Pressure (Pa)	1.008e+05	1.019e+05	1.012e+05	205.9
Volume Velocity Magnitude (m/s)	1.049	40.73	20.99	11.67

meshes (Krakos et al., 2025; Economon et al., 2024). Convective fluxes use flux-difference splitting with low-speed preconditioning for the steady RANS configuration applied to SHIFT-Wing, and a hybrid upwind-centered low-dissipation scheme for the transient DDES configuration applied to SHIFT-SUV. Second-order spatial accuracy is achieved via limited reconstruction of left and right face states from cell-centered values and gradients, using proprietary limiters resilient to geometry-induced mesh artifacts (such as large jumps in control-volume size). All computationally intensive operations execute on the GPU with double-precision arithmetic throughout; host-device transfers are limited to control-flow signaling. Steady solutions use implicit pseudo-time marching with adaptive CFL.

C.2 SHIFT-SUV: MESH STRATEGY

Computational meshes for each SHIFT-SUV variant are generated independently in ANSA (BETA-CAE Systems), producing hex-dominant cores with polyhedral transitions and prismatic boundary-layer stacks at the wall. A grid-refinement study against the FKFS quarter-scale validation case (Section C.5) identified a target of ~ 45M cells per variant as the operating point that balances near-body and wake resolution against per-sample computational cost. Figure A.1 shows the refinement zones used to capture both attached boundary-layer features on the vehicle surface and the unsteady wake structure downstream.

Table A.3: SHIFT-Wing Mach-0.5 (1975 cases): target quantities. Mean and Std pooled over all values. Low/High are median per-case p01/p99 for fields (to suppress single-cell outliers), min/max across cases otherwise.

Quantity	Low	High	Mean	Std
C_D	0.01159	0.05795	0.01999	0.006452
C_L	-0.1459	0.8312	0.3968	0.1882
C_M	-0.312	0.08169	-0.124	0.06626
C_f	0.008211	0.01047	0.009788	0.0002999
Surface Pressure (Pa)	1.913e+04	2.65e+04	2.339e+04	1362
Wall Shear Stress Magnitude(N/m^2)	2.386	30.56	10.68	5.316
Surface cells	3.01M	3.43M	3.25M	4.788e+04
Surface points	1.50M	1.72M	1.62M	2.394e+04
Volume cells	31.92M	33.09M	32.30M	2.319e+05
Volume points	5.87M	6.10M	5.97M	4.348e+04
Volume Pressure (Pa)	1.936e+04	2.654e+04	2.346e+04	1372
Volume Velocity Magnitude (m/s)	4.929	165.6	60.86	48.9

Table A.4: SHIFT-Wing mach-0.85 (1966 cases): target quantities. Mean and Std pooled over all values. Low/High are median per-case p01/p99 for fields (to suppress single-cell outliers), min/max across cases otherwise.

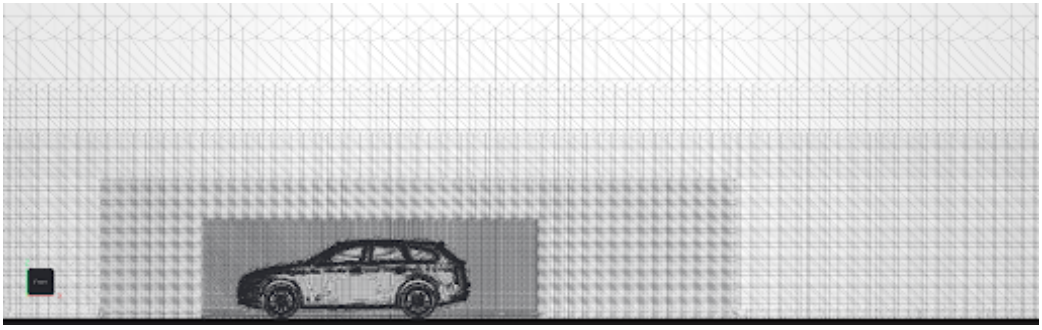
Quantity	Low	High	Mean	Std
C_D	0.01171	0.1314	0.03593	0.02274
C_L	-0.2405	0.8112	0.4631	0.2067
C_M	-0.3697	0.1173	-0.1674	0.08144
C_f	0.006849	0.009112	0.008199	0.0004089
Surface Pressure (Pa)	1.194e+04	3.225e+04	2.234e+04	3787
Wall Shear Stress Magnitude (N/m^2)	5.399	61.34	26.58	10.98
Surface cells	2.94M	3.48M	3.28M	1.255e+05
Surface points	1.47M	1.74M	1.64M	6.273e+04
Volume cells	33.02M	33.36M	33.18M	4.25e+04
Volume points	6.03M	6.18M	6.12M	3.105e+04
Volume Pressure (Pa)	1.2e+04	3.238e+04	2.25e+04	3767
Velocity Magnitude (m/s)	10.04	322.5	113.3	83.55

C.3 SHIFT-SUV: TURBULENCE MODELING

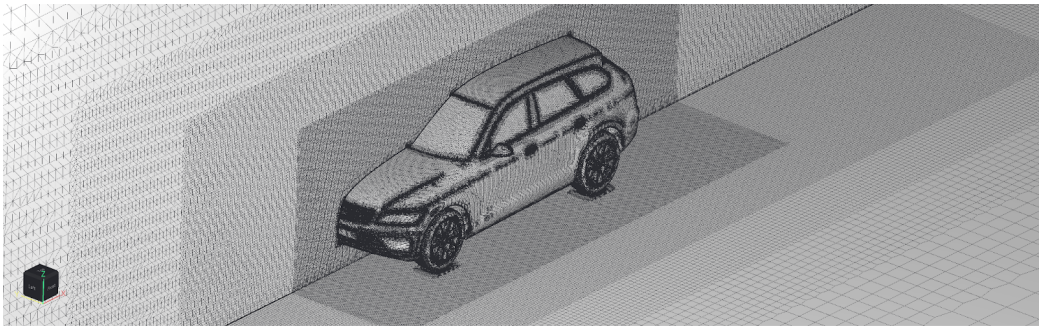
The transient regions of every SHIFT-SUV simulation use Delayed Detached Eddy Simulation (DDES) with a Spalart–Allmaras backbone (Spalart et al., 2006), configured with a shear-layer-adapted length scale and a vortex-tilting-measure (VTM) grey-area mitigation that accelerates the RANS-to-LES transition in shear-dominant regions. An advanced shielding function suppresses modeled-stress depletion, a known DDES failure mode in which the LES branch activates inside the attached boundary layer and produces artificially early flow separation. The hybrid upwind-centered convective scheme described in Section C.1 provides the low dissipation in resolved-LES zones required to support the DDES formulation.

C.4 SHIFT-SUV: DEFORMATION-CAGE PARAMETERIZATION

For the SHIFT-SUV dataset, the geometric variants are produced leveraging morphing boxes, a structured lattice free form deformation approach available in the ANSA software of BETA-CAE (Panagos & Drivakos), defined around the AeroSUV reference, as exhibited in the schematic in Figure A.2. The sixteen morphing parameters and their target ranges are listed in Table A.1; the same parameters are visualized in Figure 2. Two properties of this parameterization deserve explicit note. First, the parameters are not mutually orthogonal: many of them share cage control points, so adjusting one parameter (for example, backlight angle) modifies surfaces also influenced by oth-



(a) Side view



(b) Isometric view

Figure A.1: Mesh-refinement zones for SHIFT-SUV, sized to resolve near-body boundary-layer dynamics and the unsteady wake downstream of the vehicle.

ers (rear tapering). Second, fixed components common to every variant, such as the wheels, tires, suspension, etc are excluded from the morph, holding these regions invariant (other than the tire exposure parameter) across the dataset and isolating the geometric variation to the body shell. The target translation distances in Table A.1 are used as inputs to the cage deformation; due to the cage's smoothness constraints, they are not enforced exactly on the resulting surface.

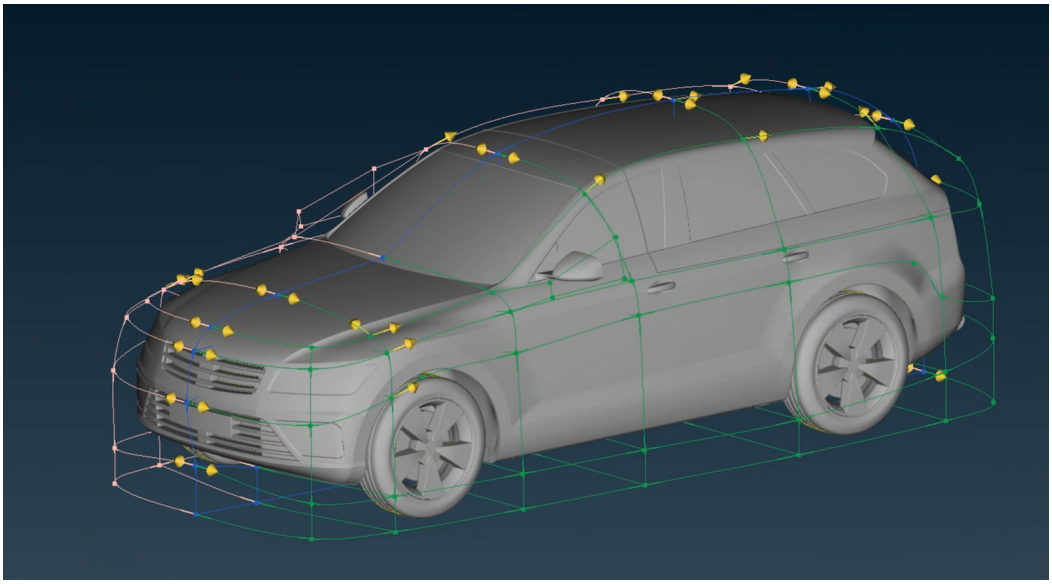


Figure A.2: Visualization of the morphing boxes and their control points defined around the Aero-SUV Estate configuration.

C.5 SHIFT-SUV: EXPERIMENTAL VALIDATION

The inflow boundary condition for the freestream for full-scale samples are shown in Table A.5, where fixed frontal area and reference length are used for computing the aerodynamic coefficients for all samples.

	U_∞	ρ_∞	μ_∞	T_∞	A_{ref}	L_{ref}
full scale	31.29 m/s	1.225 kg/m^3	1.81e^{-5} Pa-s	293.1 K	2.464 m^2	2.786 m
quarter scale	50.0 m/s	1.225 kg/m^3	1.81e^{-5} Pa-s	293.1 K	0.154 m^2	0.696 m

Table A.5: AeroSUV freestream boundary conditions.

Quantitative validation for the SHIFT-SUV CFD data was established in multiple campaigns. Simulation setup was initially validated against quarter-scale wind-tunnel measurements collected at FKFS using the AeroSUV reference geometry (Zhang, 2021). The limited experimental data available provides integrated force measurements and the delta in forces between configurations, which are used to evaluate mesh size and boundary condition choices for quarter-scale simulations.

Due to the scarce experimental data publicly available for AeroSUV, additional validation of the solver is based on our participation in the AutoCFD4 workshop (Economon et al., 2024; Hupertz et al., 2022), an industry benchmarking exercise for CFD solvers. Extensive experimental data comparisons have been performed on the DrivAer, an automotive sedan model from which the AeroSUV is derived (Zhang et al., 2019) (effectively a "taller" version of the same vehicle platform). The DrivAer data represents a full-scale vehicle, thus validating the solver for the full scale dataset featured in this manuscript.

The results of the integrated force measurements are reported in Table A.6, demonstrating a commensurate agreement between the solver and the experimental results with other participants in the workshop. Point-wise comparisons with experimental measurements in two different configurations (change in front tire windshield design) are provided in Figure A.3.

	C_D	C_L	C_{LF}	C_{LR}
Our Submission for 2a	0.2768	0.0276	-0.06948	0.097209
Our Submission for 2b	0.2669	0.0352	-0.06212	0.097355
Our Submission's Δ b-a	-0.0099	+0.0076	+0.0074	+0.00015
Experiment Δ b-a	-0.0122	-0.0056	+0.0038	-0.0094

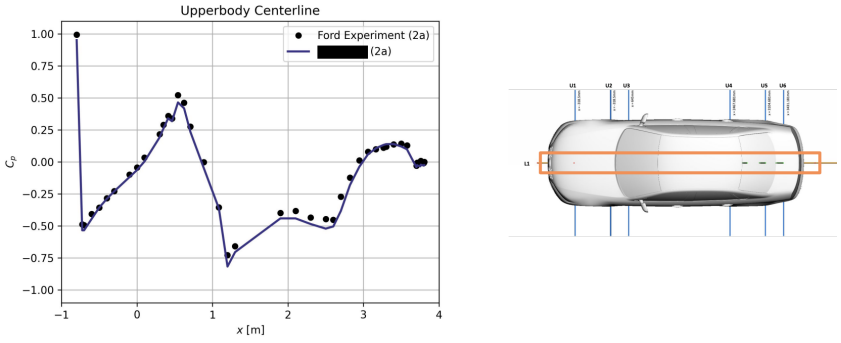
Table A.6: Comparison of aerodynamic coefficients to experiment for the DrivAer. Cases 2a and 2b refer to automotive variants without and with a front wheel air deflector.

C.6 SHIFT-WING: PARAMETRIC MODEL CONSTRUCTION

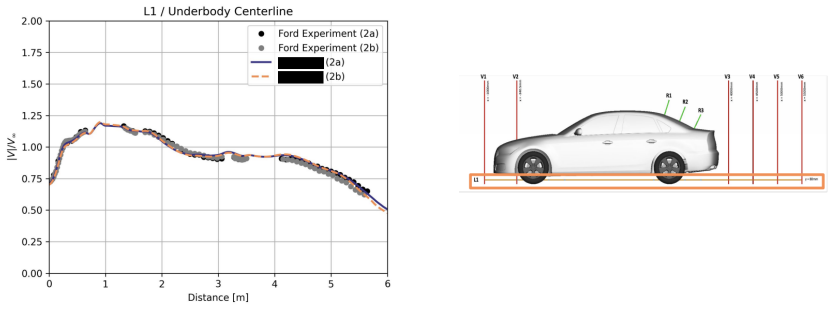
The SHIFT-Wing geometry is constructed in OnShape by reverse-engineering a parametric model from the high-speed Common Research Model (CRM) CAD distributed to AIAA Drag Prediction Workshop participants (Tinoco et al., 2018). The original CRM is described by 21 spanwise airfoil profiles (Vassberg et al., 2008), but tabulated profile data is not publicly available, requiring reconstruction from the CAD itself. The wing, fuselage and wing-body fairing are each constructed as a separate parameterized solid body and combined into the final assembly via Boolean union.

For the wing, six airfoil profiles are extracted at constant-span planes nearly uniformly distributed in 20% increments between root and tip, with the Yahudi break (37% span) substituted for the 40%-span profile. Each extracted profile is normalized to unit chord at zero twist, then perturbed via five parameters per profile (chord scaling, twist about the quarter-chord, and three-dimensional translation). Surfaces are produced by lofting the perturbed profiles in two trapezoidal sections (root-to-Yahudi and Yahudi-to-tip) and capping the result with a separately constructed rounded wing tip.

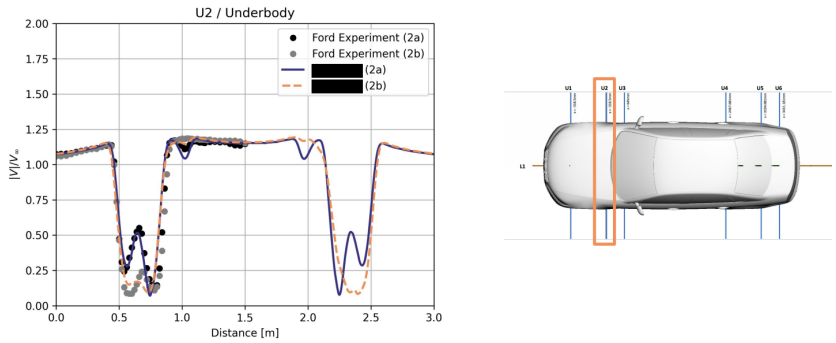
Macroscopic "wing-designer" parameters such as aspect ratio, sweep, taper ratio, and twist distribution are used to compute the per-profile parameters. Only these parameters, shown in Figure 4 and



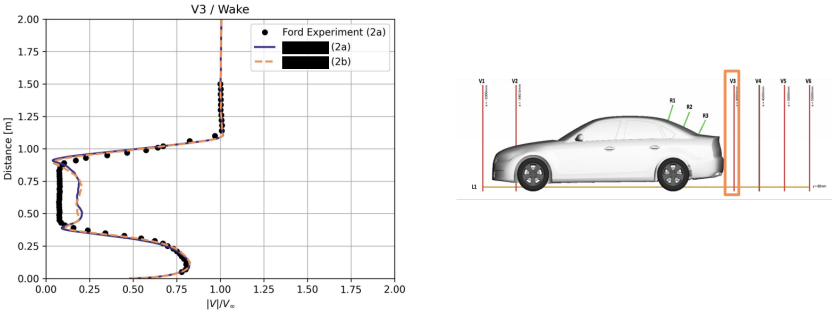
(a) C_P on the upperbody centerline



(b) Velocity magnitude underneath the vehicle



(c) Velocity magnitude spanwise distribution



(d) Velocity magnitude in the near wake

Figure A.3: Predictions of time-averaged fields compared to experimental results for the DrivAer. *Note: the submission name has been redacted to maintain anonymity. It shall be un-redacted in the camera ready version.*

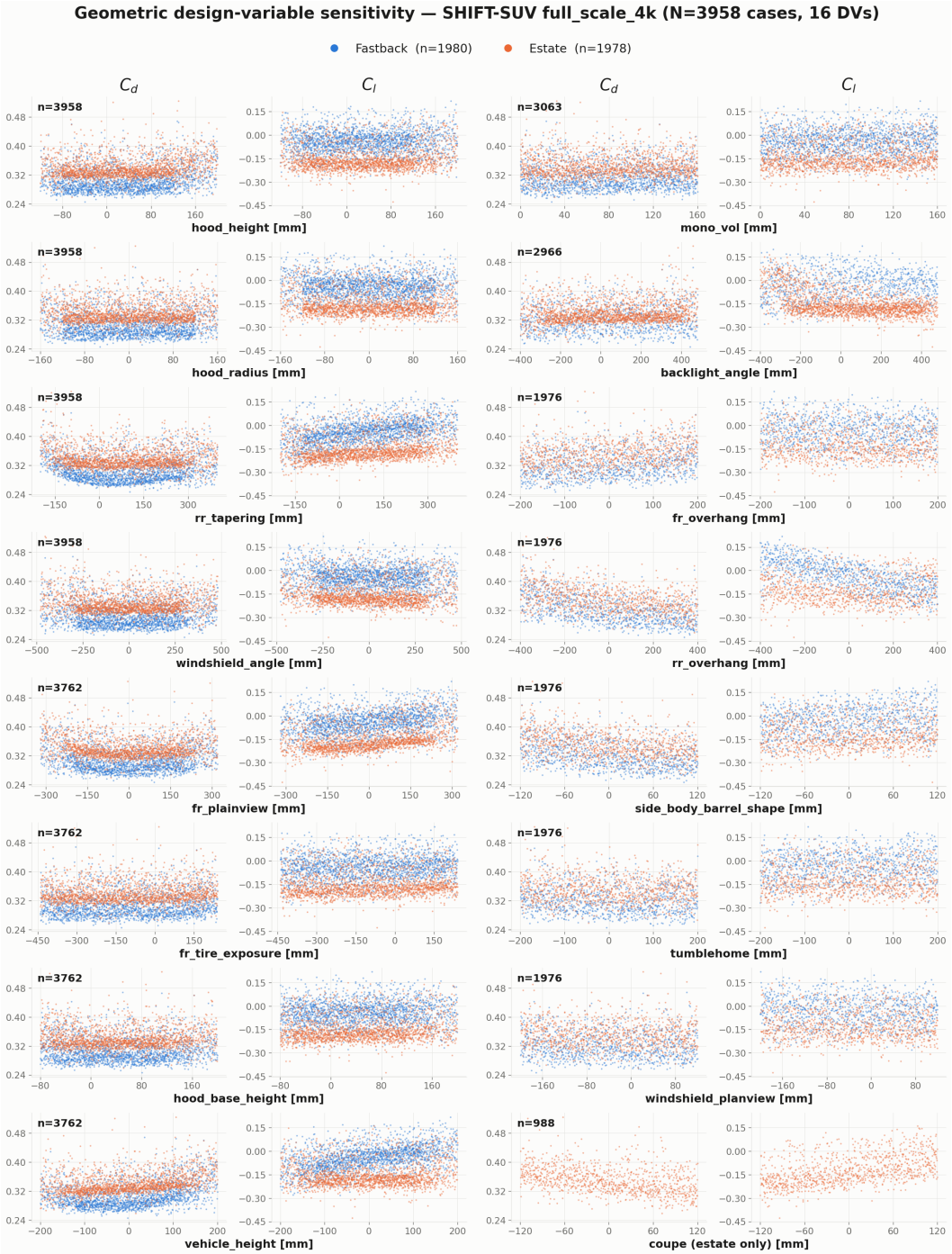


Figure A.4: Scatter plots depicting the variation of lift (C_l) and drag (C_d) coefficients against specific geometry design parameters. The design parameters are labeled in Figure 2 of the main text. In the notations, "n=xxxx" refers to number of designs where the specific design parameter is varied (active cases from Table A.1). For other designs, the specific design parameter is maintained at the baseline value.

described in Table 3, are exposed directly to the user; ensuring that geometric realizations remain consistent with how aerospace designers reason about planform design.

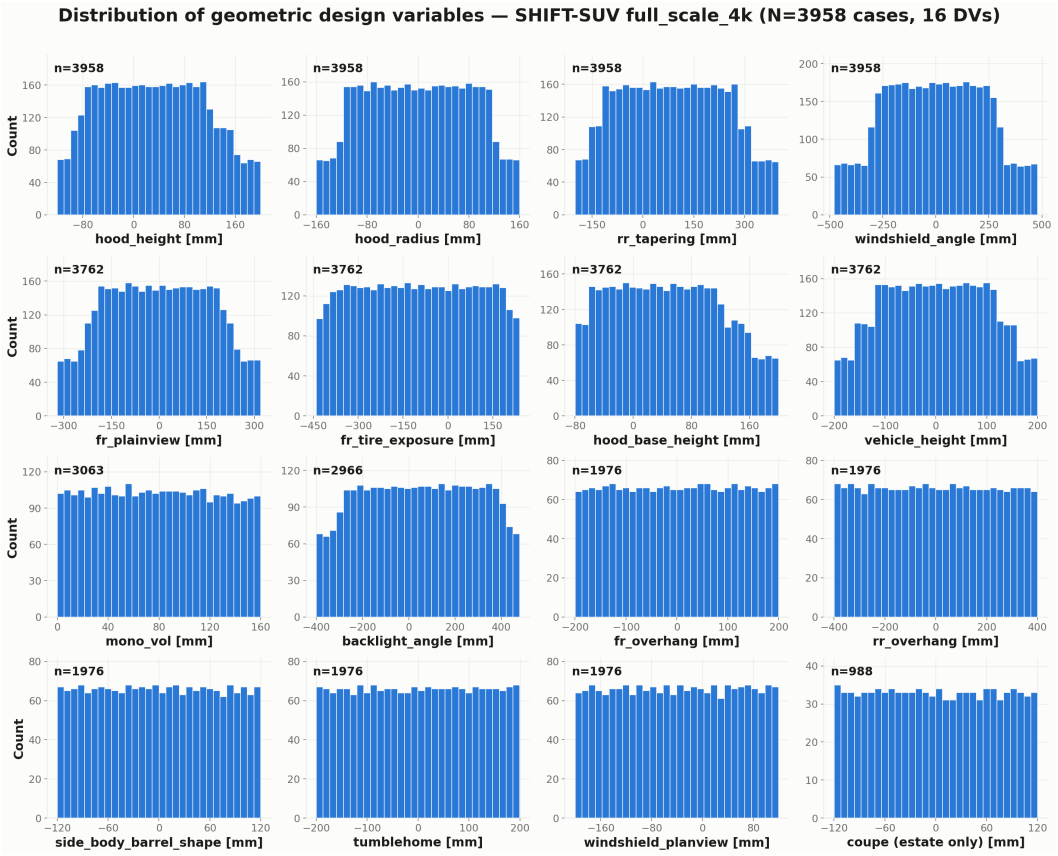
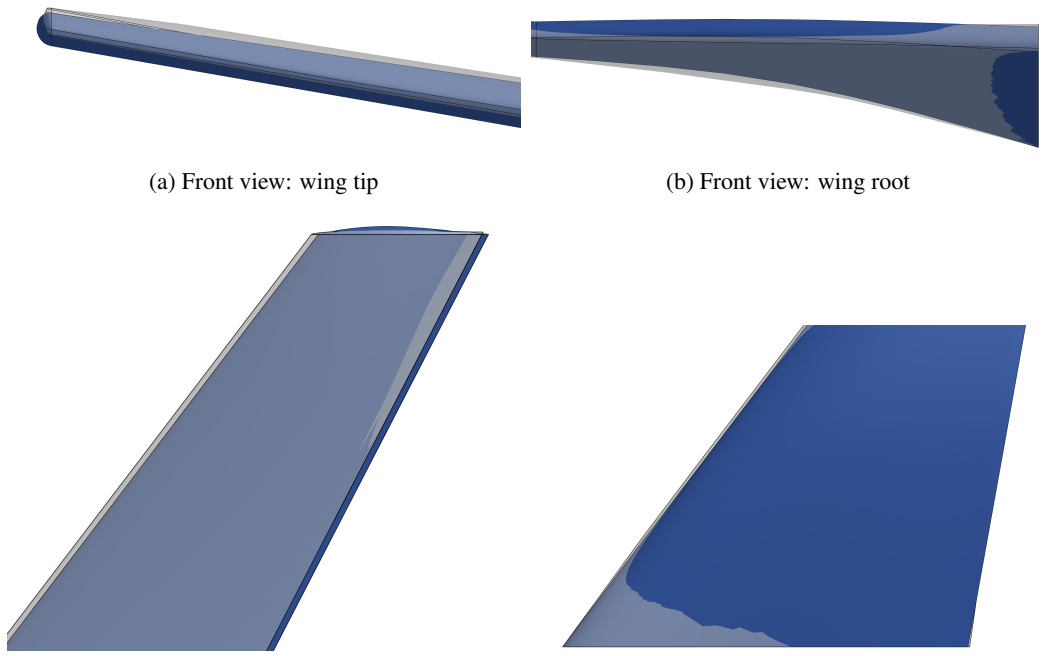


Figure A.5: Histogram depicting sample counts based on different geometry design parameters. The design parameters are labeled in Figure 2 of the main text. In the notations, "n=xxxx" refers to number of designs where the specific design parameter is varied (active cases from Table A.1). For other designs, the specific design parameter is maintained at the baseline value.

The fuselage is reduced to a parameterized cylindrical tube with retained nose and empennage geometry (uniformly scaled with fuselage diameter). The wing-body fairing is reconstructed as a new solid built from lofted spline profiles, with its center section sized parametrically to encapsulate the local wing chord and twist at the wing-body juncture. Other approaches in the literature have used surface or mesh morphing to vary the CRM (uCRM (Brooks et al., 2018; 2017) for aerostructural studies; control-surface variants for surrogate modeling (Bekemeyer et al., 2025)); the OnShape parametric route adopted here trades morph flexibility for a clean separation of the design parameters that defines the SHIFT-Wing design space.

C.7 SHIFT-WING: IMPERFECT RECONSTRUCTION

The reverse-engineered SHIFT-Wing geometry is acknowledged as an imperfect representation of the reference CRM CAD. Two sources of deviation are notable. First, the spanwise profiles are extracted on constant-span planes, which imperfectly recovers profiles from a CAD that incorporates aeroelastic deflection at the workshop-prescribed angle of attack of 2.75°. Second, the parameterized wing tip and re-lofted fairing match the reference shape closely but not exactly. Figure A.6 visualizes the residual deviation between the parameterized model at default parameter values and the reference CAD. We disclose this explicitly because the "Reference" parameter row in Table 3 should be interpreted as the closest reproduction within our parametric family rather than as an exact restatement of the original CRM.



(a) Front view: wing tip

(b) Front view: wing root

(c) Top view: wing tip

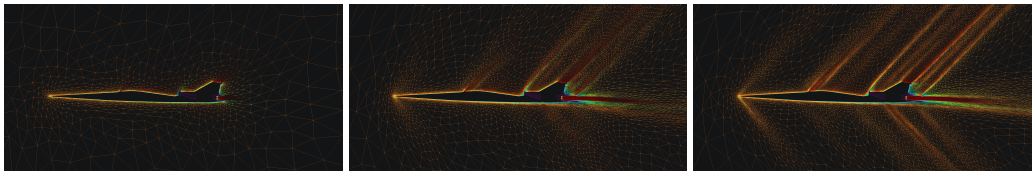
(d) Top view: wing root

Figure A.6: Residual deviation between the SHIFT-Wing parameterized model at default parameter values (blue) and the reference CRM CAD (transparent gray).

C.8 SHIFT-WING: SOLUTION-ADAPTIVE MESHING

Each SHIFT-Wing simulation produces its mesh through mesh adaptation (Krakos et al., 2025), an anisotropic solution-adaptive technology that generates a sequence of mesh-and-solution pairs, with each successive mesh refined locally and anisotropically along the previous solution's gradients via a metric-based formulation. For the transonic flow regime targeted by SHIFT-Wing, this yields meshes that are densely packed across shock structures and along leading-edge stagnation lines while remaining coarse far from the wing, producing accurate shock capture without per-sample manual mesh tuning. This is exemplified in both Figures 5 and A.7. A practical consequence is that

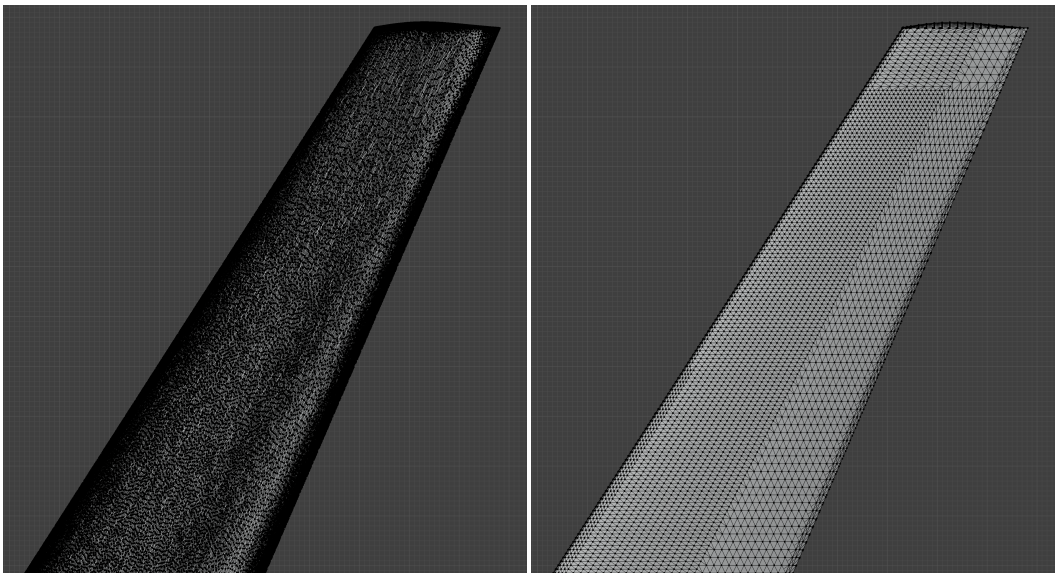
the surface-mesh node distribution varies substantially across samples even for similar geometries, since adapted nodes follow the solution rather than the geometry alone—an input distribution shift that surrogate models trained on SHIFT-Wing must accommodate. To support reproducibility, we note that this mesh adaptation framework exposes only one user-tunable parameter beyond the initial geometry: the target adapted-mesh cell count; all other adaptation behavior (refinement metric, gradation, termination) is automatic.



(a) initial solution (b) mid-sequence solution (c) final adapted solution

Figure A.7: Mesh adaptation visualized: representative mesh and solutions on the centerplane for a high-speed supersonic transport concept vehicle (NASA C608). The sequence depicts how the graph size (number of elements) and spatial distribution (node density and alignment) is driven by flow features to properly capture small scale nonlinearities such as shock waves and shear layers.

This is discussed in the context of model training and concern over the solution adapted mesh providing too much guidance on the final solution. For this reason, each case in the dataset includes an isotropic tesselation of each geometry produced by the generating CAD software, which is completely unaware of the flow solution, shown in Figure A.8.



(a) Solution-adapted: 3M nodes (b) Uniform: 200k nodes

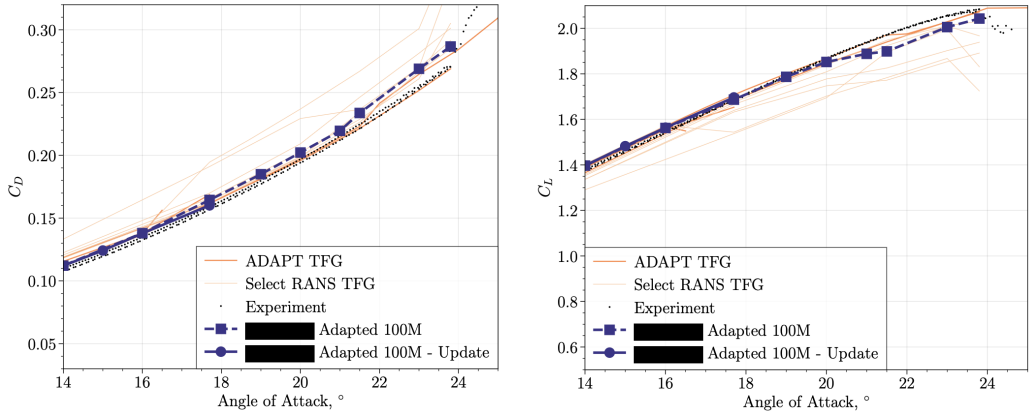
Figure A.8: Surface discretizations node distribution for a Mach 0.85 sample of SHIFT-Wing. Note the geometries of the wing are identical, simply the node distributions (their local densities) are different.

C.9 SHIFT-WING: EXPERIMENTAL VALIDATION

The mesh adaptation technology used here has been independently exercised at the 5th AIAA High Lift Prediction Workshop (Krakos et al., 2025). Validation is performed on the CRM-HL Wing-Body configuration, a CRM geometry related to the one underpinning SHIFT-Wing, and validation against experimental data on increasingly complex CRM-HL configurations including wing-body-empennage, wing-body with slats, and the full landing configuration with slat brackets. The

workshop exercised the solver beyond the flight conditions explored in this dataset with regards to the high-angle-of-attack regime; convergence and accuracy results reported in (Krakos et al., 2025) therefore provide an external, regime-distinct validation of the meshing technology applied to SHIFT-Wing.

Convergence of the solver using mesh adaptation was demonstrated on a variety of the workshop cases, one example is provided in Table 2. At relatively small mesh sizes the integrated lift, drag and moment coefficients have already converged, demonstrating the computational efficiency of the approach. Comparison to experimental results for the predicted forces across a range of angles of attack are provided in Figure A.9, highlighting the strong agreement with experiment. Note that SHIFT-Wing operates at a lower angle of attack regime than available in these plots, in a regime where the RANS modeling approach is already well trusted in industry.



(a) Drag Coefficient

(b) Lift Coefficient

Figure A.9: Comparison of integrated forces computed using mesh adaptation for Case 2.2 of the HLPW at varying angles of attack. *Note: the submission name has been redacted to maintain anonymity. It shall be un-redacted in the camera ready version.*

D Datasheets for Datasets and RAI Metadata

We provide a Datasheets for Datasets (Gebru et al., 2021) questionnaire for both SHIFT-SUV and SHIFT-Wing, structured around the seven standard headings of motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, and maintenance. The full questionnaire is released in the accompanying repository; its Responsible AI (RAI) content is also released in machine-readable form in the Croissant metadata files, one per dataset.

Because CFD data is generated rather than collected from human subjects or annotators, several Croissant Responsible AI (RAI) fields do not apply directly. We add custom fields to record provenance information that is meaningful for physical-collection validity: simulation-software version, solver convergence criteria, mesh-generation tool and parameters, validation reference (where applicable), and boundary conditions in machine-readable form. These fields are recorded as schema.org additional properties in each Croissant metadata file.

E Reproducibility Statement

The flow solver used to generate the ground-truth simulations is described in (Economon et al., 2024; Krakos et al., 2025). While the proprietary GPU-native flow solver used to generate the ground-truth data is not open-sourced, the released datasets include the complete, converged high-fidelity fields (features and targets) required to fully reproduce the neural surrogate evaluations presented in this benchmark.

F Author Statement and Data License Confirmation

The authors take full responsibility for any violation of rights arising from the release of these datasets. SHIFT-SUV and SHIFT-Wing are both released under the Creative Commons Attribution–NonCommercial 4.0 International (CC-BY-NC 4.0) license.

To ensure responsible data stewardship and mitigate the inherent dual-use risks associated with high-fidelity aerodynamic geometries (e.g., potential overlap with UAV or high-speed projectile design), access to the full dataset is gated behind a manual-approval credentialing process to track proliferation. During the double-blind review, access is provided anonymously as described in Section A.

The reference geometries on which the datasets are based, the FKFS AeroSUV platform (Zhang et al., 2019) and the NASA Common Research Model (CRM) (NASA Common Research Model Project), are themselves open-source publications. The CRM is used in accordance with its original license, and our use of AeroSUV was granted a commercial use exception by the owners. No personal or sensitive data is contained in either dataset; both consist of computational fluid dynamics simulation outputs over publicly disclosed geometries and synthetic parametric variants thereof.

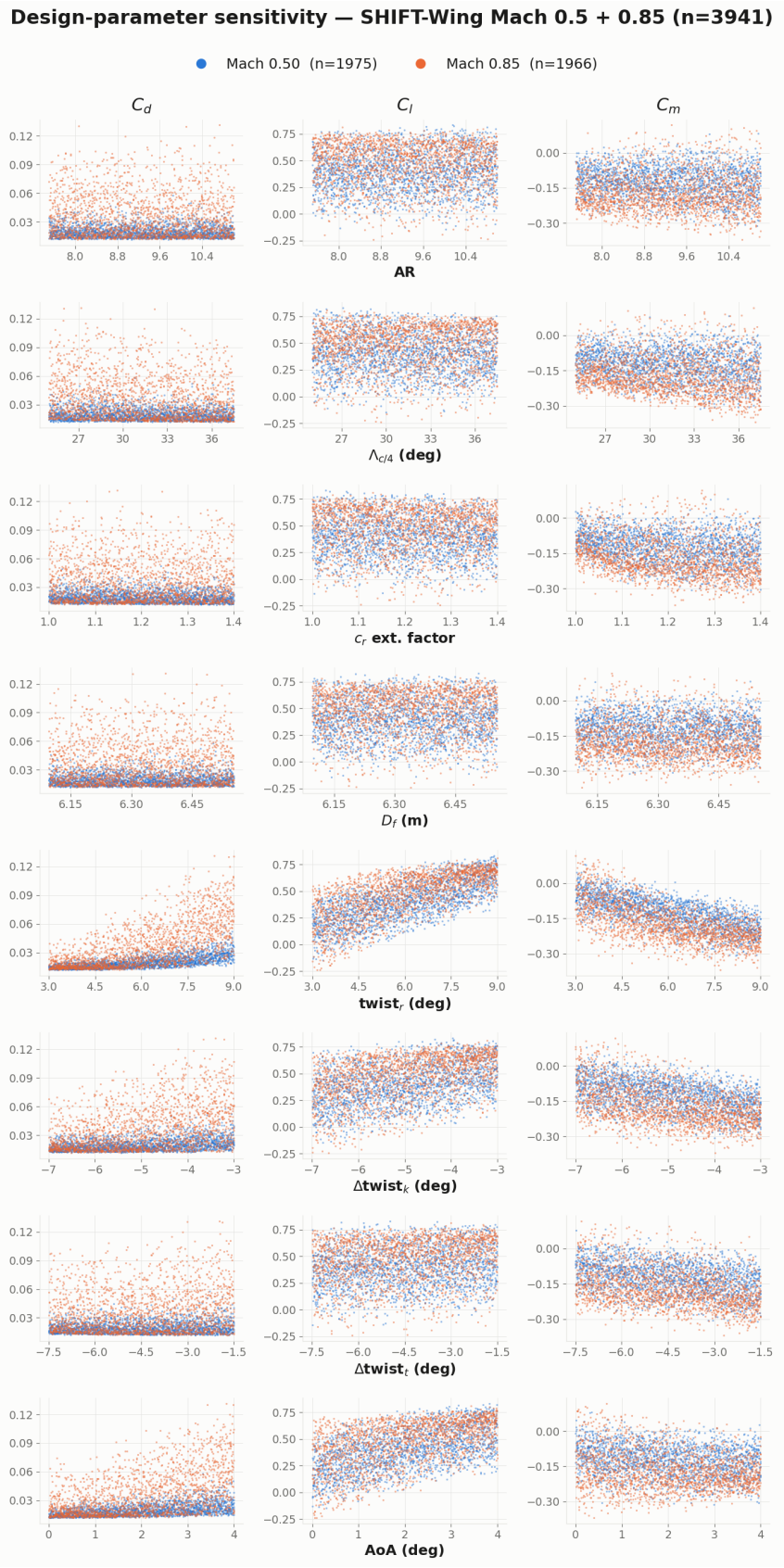


Figure A.10: Scatter plots depicting the variation of drag, lift, and moment coefficients against geometry parameters, with color indicating the flow speed.

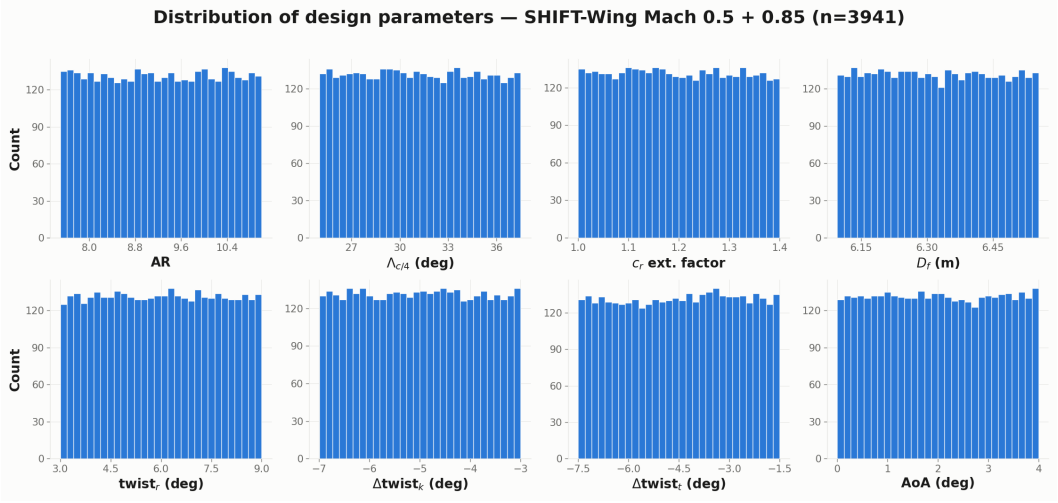


Figure A.11: Histogram depicting sample counts based on different geometry design parameters. The geometry parameters are labeled in Figure 5 of the main text. The symbols and ranges for all the parameters referenced above are defined in Table 3 of the main text.