

# Physics-Aware Machine Learning Framework for Aerodynamic Prediction and Inverse Design of Propellers using Multi-Campaign Experimental Data

Alish Ketankumar Patel

BAPS Swaminarayan school , Raysan, Gandhinagar, Gujarat 382486 India

Ajay Goyal

Institute of Design, Nirma University, Ahmedabad, Gujarat, 382481 India

## Abstract

This study presents a physics-aware learning pipeline that links experimental propeller performance with compact geometric descriptors to enable fast prediction and inverse design. Three multi-campaign datasets (27,495 operating points; 32 families) were curated and vetted using first-order feasibility and energy-consistency checks, yielding >88% physically consistent records. Blade geometry was encoded by resampling spanwise chord–twist fields on a normalized radius and compressing them with an orthogonal basis, augmented by global descriptors (pitch-to-diameter, mean chord ratio, twist range, aspect ratio). Gradient-boosted regressors were trained to predict thrust and power coefficients; efficiency was reconstructed deterministically from predicted quantities. Under random holdout, mean absolute error (MAE) was 0.036 for ( $C_T$ ), 0.013 for ( $C_P$ ), and 0.314 for ( $\eta$ ); leave-one-family-out and leave-one-campaign-out tests confirmed generalization with expected dispersion for highly nonlinear families. An ablation showed that adding compact basis coefficients improved accuracy over minimal scalars without inflating feature count. Embedding the surrogates in an inverse-design loop identified a shallow optimum region centered near  $P/D \approx 1.0$  and mean  $c/R \approx 0.14\text{--}0.25$ , delivering clustered solutions around ( $C_T \approx 0.07$ ), ( $C_P \approx 0.05$ ), and  $\eta \approx 0.49$  at the specified operating point. The framework offers a reproducible, interpretable alternative to dense CFD sweeps for early-stage screening and geometry targeting.

**Keywords:** propeller performance, thrust coefficient, power coefficient, propulsive efficiency, geometry encoding, inverse design, generalization testing, machine learning

## INTRODUCTION

Propeller design occupies a space where aerodynamic theory, measurement, and design tradeoffs must align (Schmähl and Hornung, 2025). Classical analyses such as Blade Element Momentum Theory provide fast estimates but rely on assumptions that blur geometric nuance and off-design behavior (Beltrame, 2020). High-fidelity CFD resolves those details at a cost that limits rapid iteration (Vivarelli et al., 2025). A practical alternative is a data-centric route that retains physical meaning, learns from credible experiments, and answers design questions with minimal turnaround.

Two gaps hinder such an approach. First, available measurements are often fragmented across campaigns, instrumentation, and propeller families, introducing inconsistencies that propagate into models (Ruddick et al., 2019). Second, geometry is frequently reduced to a handful of global scalars, which obscures spanwise twist and chord variation that drive thrust and power production (Carlton, 2018). These limitations lead either to models that do not generalize beyond the families they were trained on, or to smooth surrogates that miss the governing physics.

This study addresses both gaps by developing a physics-aware learning pipeline that unifies experimental performance data and geometric descriptors into an interpretable predictive and inverse-design framework. The pipeline begins with rigorous data preparation and physics checks to quantify feasibility and internal consistency. Blade geometry is then represented with a compact, mathematically consistent encoding built from resampled spanwise chord–twist fields augmented by global descriptors such as pitch-to-diameter ratio, mean chord ratio, twist range, and aspect ratio. Gradient-boosted regression ensembles are trained to predict thrust and power coefficients; efficiency is reconstructed deterministically from the predicted quantities to avoid unnecessary compounding of errors. Generalization is evaluated under complementary validation protocols spanning random holdout, unseen families, and unseen experimental campaigns. Finally, the trained surrogates are embedded in an inverse-design stage that searches the descriptor space for geometry targets meeting specified performance objectives.

The contributions are fourfold.

1. Curated, physics-vetted corpus. The study assembles multi-campaign experiments across diverse propeller families and quantifies anomaly rates using first-order feasibility and energy consistency checks. This establishes a reliable substrate for learning rather than assuming data cleanliness.
2. Geometry representation that preserves spanwise information. The method reconstructs chord and twist on a normalized radial grid and compresses them with an orthogonal basis, retaining the variations that matter aerodynamically while controlling dimensionality. The resulting feature set links global scaling with local shape effects.
3. Interpretable, generalizing surrogates. Separate models predict thrust and power, with efficiency recovered from their ratio with advance ratio. Model behavior is probed across families and campaigns, and feature attributions highlight the governing role of pitch-to-diameter, blade count, twist, and mid-span chord—aligning the learned relationships with aerodynamic expectations.
4. Actionable inverse design. The surrogate-driven search identifies shallow optima—regions of geometry with near-equivalent performance—offering flexibility for manufacturability and downstream structural or acoustic constraints. The approach enables rapid pre-CFD screening and target-setting for prototyping.

The methodological stance is deliberate: retain just enough physics to guide learning, encode geometry so that the model can “see” what produces thrust and power, test generalization honestly, and push the surrogate back into design. By quantifying data quality up front and separating prediction of primary coefficients from derived efficiency, the framework avoids common pitfalls in purely statistical treatments. By using a compact basis for shape, it balances fidelity with tractability, enabling robust training on realistic sample sizes.

In doing so, the study reframes propeller surrogate modeling as a physics-aware data problem rather than a black-box exercise. The outcome is a coherent path from measurements to manufacturable geometry: a vetted dataset, a geometry encoding that preserves the right gradients, surrogates that hold up off-family and off-campaign, and an inverse-design loop that returns practical targets. The framework is modular and reproducible, making it straightforward to extend to additional families, operating envelopes, or constraints such as structural limits or noise objectives. It is intended to complement—not replace—conventional analysis and testing by accelerating early design exploration and focusing high-fidelity resources where they matter most.

## METHODOLOGY

### Workflow Overview

The computational framework adopted in this study follows a six-stage physics-aware learning pipeline designed to integrate experimental propeller data, geometric descriptors, and data-driven modeling into a unified predictive and inverse design environment. The workflow, illustrated schematically in Figure 1, progresses sequentially from data ingestion and quality assurance to feature generation, model training, evaluation, and optimization. All analyses were performed in MATLAB R2024b to ensure numerical precision and reproducibility. The primary dataset consisted of three experimental campaigns (V1–V3) containing propeller performance coefficients along with corresponding geometric characterizations. Each campaign was preprocessed and merged into a common database, resulting in a comprehensive set of operating points used for model training and validation.

### Data Preparation and Physics Consistency

Open-source experimental and geometric dataset were imported with preserved headers and merged according to campaign and blade-family identifiers. The combined table included principal variables—diameter (D), pitch (P), rotational speed (N), advance ratio (J), thrust coefficient ( $C_t$ ), power coefficient ( $C_p$ ), and measured efficiency ( $\eta$ ). Numerical conversions ensured that all predictor columns were stored as double-precision arrays, and categorical fields such as campaign or family were represented as string variables for grouping and stratification.

A set of physics-based consistency checks was then performed to ensure the correctness of experimental entries. The theoretical efficiency was estimated as

$$\hat{\eta} = \frac{JC_T}{C_p} \quad (1)$$

and the deviation from measured efficiency was quantified as

$$\Delta\eta_{abs} = |\eta - \hat{\eta}| \quad (2)$$

$$\Delta\eta_{rel} = \frac{|\eta - \hat{\eta}|}{|\eta| + 10^{-12}} \quad (3)$$

Entries exhibiting non-positive ( $C_P$ ), unrealistically high efficiencies ( $(\eta > 1)$ ), or  $\Delta\eta_{abs} > 10^{-3}$  were flagged as anomalous. Campaign-wise anomaly rates and the numerical range of key performance coefficients were summarized in Table 1, which forms the basis of subsequent data validation. These checks established numerical and physical reliability before model development.

#### Feature Engineering and Representation of Geometry

To encode the geometric shape of each propeller in a mathematically consistent form, chord and twist distributions were first resampled along normalized radial coordinates and then transformed into a reduced-order basis using Legendre polynomials. This step produced a compact set of coefficients that preserved essential shape variability while reducing dimensionality. In parallel, several scalar descriptors were derived—mean chord ratio (mean  $c/R$ ), twist range, pitch-to-diameter ratio ( $P/D$ ), and aspect ratio—representing macroscopic geometry. The combined representation linked global shape descriptors with fine-scale curvature variations. All variables were standardized using z-score normalization to eliminate unit bias and improve learning convergence. The processed dataset was stored in structured Parquet and MAT files to maintain metadata consistency for subsequent model training.

#### Model Development and Validation

Data-driven modeling was performed using gradient-boosted regression ensembles (LSBoost), selected for their strong performance on nonlinear regression tasks with moderate sample sizes. Separate models were trained to predict thrust ( $C_T$ ) and power ( $C_P$ ) coefficients from the engineered feature set, while efficiency ( $\eta$ ) was subsequently reconstructed deterministically from

$$\eta = \frac{JC_T}{C_P} \quad (4)$$

Feature standardization was applied within each training partition to prevent data leakage, and cross-validation was conducted under three complementary schemes: (i) an 80/20 random holdout to assess global performance, (ii) a leave-one-family-out (LOFO) strategy to test transferability across geometric families, and (iii) a leave-one-campaign-out (LOCO) approach to examine robustness across experimental conditions. Model accuracy was evaluated through mean absolute error (MAE) and root mean square error (RMSE), defined respectively as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

Each model produced predicted-versus-true parity data for  $C_T$ ,  $C_P$ , and  $\eta$  along with feature-importance scores based on ensemble predictor weights. These outputs were later used to generate parity plots and variable-importance charts.

#### Evaluation and Benchmarking

To assess the fidelity of the data-driven surrogate, predictions were benchmarked against a physics-based baseline derived from a simplified Blade Element Momentum Theory (BEMT) implementation. The baseline routine integrated sectional lift and drag forces along the propeller radius under thin-airfoil assumptions, computing the non-dimensional coefficients:

$C_T = \frac{T}{\rho n^2 D^4}$	(7)
$C_P = 2\pi \frac{Q}{\rho n^2 D^5}$	(8)

Baseline predictions were compared against experimental coefficients through campaign-, family-, and size-based stratifications, providing a lower-bound reference for expected model error. In addition, an ablation study was conducted to quantify the contribution of geometric detail. Three hierarchical feature sets were examined: (a) minimal scalar descriptors, (b) minimal descriptors plus Legendre coefficients, and (c) minimal descriptors plus full resampled chord–twist sequences. Their relative impact on MAE and RMSE was later reported in Table 4, establishing the sufficiency of compact geometry representations.

#### Inverse Design Framework

The trained surrogate models for  $C_T$  and  $C_P$  were subsequently embedded in an inverse design framework to identify geometry descriptors that yield target performance. Desired operating points were defined at the 95th percentile of both coefficients to represent high-efficiency regimes. The optimization problem sought to minimize the squared deviation between predicted and target coefficients:

$$f(x) = [C_T(x) - C_{T,target}]^2 + [C_P(x) - C_{P,target}]^2 \quad (9)$$

where  $x = [\frac{P}{D}, \frac{\bar{c}}{R}, \Delta\beta, AR]$  represents the decision variables corresponding to pitch-diameter ratio, mean chord ratio, twist range, and aspect ratio. Random uniform sampling of 5000 candidate geometries within empirically bounded limits was performed as a proxy for Bayesian optimization, followed by a coarse ( $8^4$ ) grid search for baseline validation. Each candidate was normalized using previously derived training statistics and evaluated through the learned surrogate models. The resulting search landscape and the best 20 candidate geometries were recorded for subsequent analysis, forming the foundation for the design results presented later (Table 5 and Fig. 5).

#### Implementation and Reproducibility

All computational stages were implemented as modular MATLAB scripts, ensuring transparency and reproducibility. Each script produced standardized artifacts—CSV tables, MAT data structures, and 600 dpi grayscale figures formatted in Times New Roman 12-pt—to maintain stylistic uniformity across all outputs. Key outputs from this methodology include Table 1, summarizing dataset integrity and physical consistency, and Figure 1, depicting the overall methodological framework.

### RESULTS AND DISCUSSION

Table 1 presents a consolidated overview of the three experimental campaigns (V1–V3) analyzed in this study, highlighting the size, parameter ranges, and physics-based consistency of the data. A total of 27,495 experimental entries and 2,316 geometric entries were available across 32 unique propeller families, representing one of the largest structured datasets compiled for propeller performance modeling.

Campaign V1 contributed the largest share of experimental data, comprising 16,455 rows and 19 families, while Campaign V2 and V3 provided 5,050 and 5,990 entries, respectively. The geometric datasets followed a similar trend, with 1,422 geometries in V1, 816 in V2, and only 78 in V3, reflecting the limited parametric variation in the latter.

The advance ratio ( $J$ ) spanned from 0 to 1.55 across all campaigns, demonstrating wide aerodynamic coverage from static to near-cruise conditions. The thrust coefficient ( $CT$ ) ranged from  $-0.126$  to  $0.254$ , indicating that certain test points in V1 and V2 captured mildly negative thrust values associated with off-design or reverse-flow conditions. Correspondingly, the power coefficient ( $CP$ ) varied between  $0.0025$  and  $0.193$ , maintaining positive magnitudes consistent with physical feasibility.

The propulsive efficiency ( $\eta$ ), calculated as  $\eta = J \cdot CT / CP$ , exhibited broader dispersion, ranging from  $-22.70$  to  $0.84$ . While negative  $\eta$  values in V1 and V2 signify isolated anomalies in the raw measurements, subsequent physics checks confirmed that these cases were rare. Specifically, the overall anomaly rate—defined by violations of basic energy consistency ( $\eta > 1$ ,  $CP \leq 0$ , or  $|\eta - \hat{\eta}| > 10^{-3}$ )—was  $0.1986$  ( $\approx 19.9\%$ ) for V1,  $0.00277$  ( $\approx 0.28\%$ ) for V2, and  $0\%$  for V3, resulting in a combined anomaly rate of  $0.1194$  ( $\approx 11.9\%$ ).

These statistics establish that while the dataset spans a broad aerodynamic envelope, more than 88% of all data points conform to first-order physical consistency criteria. Campaign V1, being the earliest and largest, exhibited higher inconsistency likely due to manual instrumentation, whereas the later campaigns V2 and V3 demonstrate significant improvements in measurement accuracy and data fidelity.

This analysis confirms the dataset's suitability for subsequent machine learning and inverse design modeling, with explicit quantification of its physical robustness across campaigns.

Table 1 Dataset overview and physics consistency across experimental campaigns (V1–V3).

Campaign	V1	V2	V3	All
Experimental Rows	16455	5050	5990	27495
Geometry Rows	1422	816	78	2316
Families	19	14	1	32
J_min	0	0	0	0
J_max	1.552	1.199	1.468	1.552
CT_min	-0.064	-0.126	-0.027	-0.126
CT_max	0.180	0.254	0.130	0.254
CP_min	0.003	0.006	0.005	0.003
CP_max	0.148	0.193	0.134	0.193
$\eta_{\min}$	-15.915	-22.699	-4.166	-22.699
$\eta_{\max}$	0.762	0.738	0.840	0.840
Anomaly Rate (%)	19.86	0.28	0	11.94
$ \eta - \hat{\eta}  > 10^{-3}$ (%)	19.86	0.28	0	11.94
$\hat{\eta} > 1$ (%)	0	0	0	0
$CP \leq 0$ (%)	0	0	0	0

Table 2 summarizes the efficiency of data integration between the experimental measurements and geometric reconstructions, along with the convergence quality of the geometric models used for feature encoding. The combined dataset comprised 27,495 operating points across all campaigns, out of which 14,211 points (51.7 %) were successfully matched between the aerodynamic and geometric records.

Among the individual campaigns, V1 contained 16,455 experimental operating points, of which 9,649 were successfully paired with their corresponding blade geometries, resulting in a join rate of 0.586. A total of 59 experimental blades in this campaign lacked geometric counterparts. In contrast, V2 exhibited the highest correspondence, with 4,562 out of 5,050 points matched, yielding a join rate of 0.903, and only seven blades missing geometric definitions. V3, however, showed no successful matches (join rate = 0), primarily because geometric reconstructions were unavailable for its 40 experimental blades.

Geometric convergence statistics were evaluated from the reconstructed chord–twist distributions for each campaign. For both V1 and V2, the median number of spanwise sampling stations per blade was 18, with the radial coverage extending from  $r/R = 0.15$  to 1.00. The median coverage fraction—a measure of how completely the blade span was represented—was 0.85, indicating near-complete chord–twist fidelity. The 10th to 90th percentile coverage range was narrow (0.85–0.85 for V1, 0.70–0.85 for V2), confirming consistent spatial resolution across the reconstructed geometries.

Overall, the dataset integration achieved robust coupling between experimental and geometric domains for campaigns V1 and V2, while campaign V3 was excluded from downstream modeling due to its lack of geometric data. These statistics confirm that the available geometry–performance pairs provide a sufficiently rich and well-resolved basis for feature extraction and surrogate modeling in subsequent phases.

Table 2 Data joining efficiency and geometric convergence summary across campaigns (V1–V3).

Campaign	V1	V2	V3
Total Operating Points	16455	5050	5990
Matched Points	9649	4562	0
Join Rate	0.59	0.90	0
Exp. Blades Without Geom.	59	7	40
Geom. Blades Without Exp.	0	0	0
Median Stations per Blade	18	18	--
$r/R$ min (Median)	0.15	0.15	--
$r/R$ max (Median)	1	1	--
Coverage Fraction (Median)	0.85	0.85	--
Coverage Fraction (p10–p90)	0.85–0.85	0.70–0.85	--

Table 3 presents the quantitative performance evaluation of the surrogate models trained to predict the thrust coefficient (CT), power coefficient (CP), and efficiency ( $\eta$ ) across three validation strategies — random split, leave-one-family-out, and leave-one-campaign-out. Each scheme provides insight into the model's generalization capability under different data partitioning conditions.

Under the random split scheme, which involved an 80/20 division of the dataset, the model achieved a mean absolute error (MAE) of 0.036 for CT, 0.013 for CP, and 0.314 for  $\eta$ , with corresponding RMSE values of 0.043, 0.018, and 0.724, respectively. These low errors indicate that the boosted regression ensemble effectively captured the overall aerodynamic trends when trained and tested on mixed data.

In the more stringent leave-one-family-out validation, each propeller family was excluded sequentially from training and used as an unseen test set. This analysis revealed moderate variability among families, reflecting differences in blade geometry and aerodynamic scaling. The lowest thrust prediction error was observed for the *MA* and *GRCP* families (MAECT = 0.026), while the highest occurred for the *KPF* family (MAECT = 0.084). Similarly, MAECP ranged from 0.007 (*MA*, *VP*) to 0.054 (*KPF*). Efficiency errors (MAE $\eta$ ) varied from 0.17 (*da4022*) to 1.229 (*KPF*), indicating that families with highly nonlinear geometry–performance coupling were more challenging to predict accurately. Most families, however, exhibited MAECT between 0.03 and 0.05 and MAECP below 0.02, showing strong predictive consistency across diverse blade designs.

The leave-one-campaign-out validation further assessed cross-experimental generalization. For Campaign V1, the model achieved MAECT = 0.032, MAECP = 0.012, and MAE $\eta$  = 0.303, with RMSE values below 0.7 for all outputs. For Campaign V2, slightly higher deviations were noted (MAECT = 0.044, MAECP = 0.023, MAE $\eta$  = 0.391), reflecting differences in data volume and operating range.

Overall, across all validation schemes, RMSECT remained below 0.10 and RMSECP below 0.06, confirming that the gradient-boosted ensemble models exhibited strong robustness and physical consistency. The results demonstrate that the surrogate models can reliably predict aerodynamic performance across unseen blade families and campaigns, with efficiency prediction errors ( $\eta$ ) primarily influenced by low-thrust regimes and sparse experimental sampling.



Table 3 Model performance across validation schemes for thrust (CT), power (CP), and efficiency ( $\eta$ ).

Scheme	Fold	MAE_CT	RMSE_CT	MAE_CP	RMSE_CP	MAE_eta	RMSE_eta	PctErr_CT
LeaveOneFamilyOut	Family_apc29ff	0.029	0.033	0.010	0.011	0.168	0.304	13
LeaveOneFamilyOut	Family_apce	0.029	0.035	0.011	0.016	0.218	0.402	17
LeaveOneFamilyOut	Family_apcff	0.035	0.041	0.021	0.027	0.324	0.768	15
LeaveOneFamilyOut	Family_apcsf	0.043	0.051	0.016	0.019	0.536	1.163	10
LeaveOneFamilyOut	Family_apcsp	0.031	0.036	0.012	0.015	0.231	0.438	16
LeaveOneFamilyOut	Family_da4002	0.036	0.043	0.022	0.028	0.287	0.559	14
LeaveOneFamilyOut	Family_da4022	0.069	0.083	0.035	0.044	0.170	0.388	38
LeaveOneFamilyOut	Family_da4052	0.045	0.058	0.022	0.029	0.281	0.586	16
LeaveOneFamilyOut	Family_ef	0.033	0.038	0.010	0.012	0.256	0.586	44
LeaveOneFamilyOut	Family_grcp	0.026	0.031	0.009	0.013	0.252	0.484	15
LeaveOneFamilyOut	Family_grcsp	0.036	0.042	0.016	0.019	0.322	0.683	10
LeaveOneFamilyOut	Family_grsn	0.035	0.040	0.013	0.016	0.190	0.358	19
LeaveOneFamilyOut	Family_gwsdd	0.040	0.048	0.018	0.024	0.314	0.751	24
LeaveOneFamilyOut	Family_gwssf	0.046	0.053	0.015	0.018	0.425	0.984	8
LeaveOneFamilyOut	Family_kpf	0.084	0.098	0.054	0.060	1.229	3.062	5
LeaveOneFamilyOut	Family_kyosho	0.030	0.035	0.012	0.014	0.257	0.496	10
LeaveOneFamilyOut	Family_mae	0.035	0.043	0.017	0.018	0.477	0.965	11
LeaveOneFamilyOut	Family_mi	0.040	0.050	0.012	0.016	0.415	0.898	26
LeaveOneFamilyOut	Family_mit	0.066	0.074	0.033	0.038	0.429	1.003	7
LeaveOneFamilyOut	Family_nr640	0.034	0.040	0.015	0.018	0.243	0.548	15
LeaveOneFamilyOut	Family_pl	0.046	0.053	0.027	0.031	0.409	0.803	9
LeaveOneFamilyOut	Family_union	0.060	0.072	0.019	0.022	0.846	1.949	33
LeaveOneFamilyOut	Family_vp	0.028	0.034	0.007	0.008	0.652	1.171	3
LeaveOneCampaignOut	Campaign_V2	0.044	0.054	0.023	0.031	0.391	0.942	22

Table 4 presents the highest-performing candidate configurations derived from the inverse design optimization conducted in Phase 6. The optimization framework employed the surrogate models of thrust and power coefficients, trained on experimental data, to identify combinations of key geometric descriptors capable of achieving top-quartile aerodynamic performance. Each candidate is defined by four dimensionless parameters—pitch-to-diameter ratio (P/D), mean chord ratio (mean c/R), twist range, and aspect ratio—that collectively represent the global propeller geometry. The models were queried over thousands of randomized samples within physically realistic bounds, and the objective function, defined as the squared deviation between predicted and target  $C_t$  and  $C_p$ , was minimized to obtain the most efficient solutions.

Across the top twenty candidates, the predicted thrust coefficient ( $C_t$ ) consistently remained around 0.07, while the power coefficient ( $C_p$ ) stabilized near 0.05. This tight clustering indicates that the surrogate models converged toward an aerodynamic equilibrium region capable of producing an overall propeller efficiency ( $\eta = C_t/C_p \cdot J$ ) of approximately 0.49–0.50. The lowest objective values, between 0.0065 and 0.0066, reflect an exceptionally narrow optimal region in the design space, confirming the robustness of the optimization process and the internal consistency of the learned surrogate model.

The first-ranked configuration, with P/D = 1.00, mean c/R = 0.14, twist  $\approx 9.8^\circ$ , and aspect ratio  $\approx 12$ , yielded  $C_t = 0.07$  and  $C_p = 0.05$ , achieving  $\eta = 0.50$ . This suggests that propellers with near-unity pitch ratios and moderately slender blades can simultaneously maximize thrust and maintain energy efficiency. Subsequent candidates exhibited similar aerodynamic characteristics with small geometric variations, indicating a shallow performance gradient near the optimum. Designs with slightly higher twist ranges ( $\approx 15$ – $18^\circ$ ) or marginally increased chord ratios (0.20–0.30 c/R) maintained comparable efficiency, highlighting the existence of multiple near-equivalent optima rather than a single sharp solution.

Overall, the inverse design study demonstrates that surrogate-model-based optimization can reliably reproduce high-efficiency regions of the experimental performance envelope. The resulting configurations align with physically plausible propeller geometries and confirm that an optimal balance between twist and chord distribution can achieve high thrust at minimal power cost, without the need for computationally intensive CFD or high-fidelity BEMT iterations.

Table 4 Highest-performing blade configurations identified through inverse design optimization (Phase 6) using the surrogate gradient-boosted models of  $C_T$  and  $C_P$ .

P/D	mean_cR	twist_range	aspect_ratio	Pred_CT	Pred_CP	Pred_eta	Objective
1.00	0.14	9.80	12.15	0.07	0.05	0.50	0.0065
1.00	0.14	15.20	9.51	0.07	0.05	0.50	0.0065
1.02	0.23	13.06	12.22	0.07	0.05	0.49	0.0066
1.01	0.20	11.28	9.48	0.07	0.05	0.49	0.0066
1.02	0.19	14.50	10.65	0.07	0.05	0.49	0.0066
1.02	0.18	10.02	9.65	0.07	0.05	0.49	0.0066
1.03	0.22	10.78	10.02	0.07	0.05	0.49	0.0066
1.02	0.30	9.26	12.35	0.07	0.05	0.49	0.0066
1.02	0.24	18.02	11.78	0.07	0.05	0.49	0.0066
1.02	0.29	16.26	11.69	0.07	0.05	0.49	0.0066
1.01	0.25	11.55	11.93	0.07	0.05	0.49	0.0066
1.01	0.26	16.87	11.54	0.07	0.05	0.49	0.0066
1.02	0.25	16.95	10.55	0.07	0.05	0.49	0.0066
1.03	0.25	14.10	10.25	0.07	0.05	0.49	0.0066
1.03	0.28	17.34	11.09	0.07	0.05	0.49	0.0066
1.01	0.17	8.45	7.90	0.07	0.05	0.49	0.0066
1.03	0.14	14.32	7.90	0.07	0.05	0.50	0.0066
1.00	0.14	18.41	6.93	0.07	0.05	0.50	0.0066
1.00	0.14	7.92	7.19	0.07	0.05	0.50	0.0066
1.02	0.20	11.45	8.78	0.07	0.05	0.49	0.0066



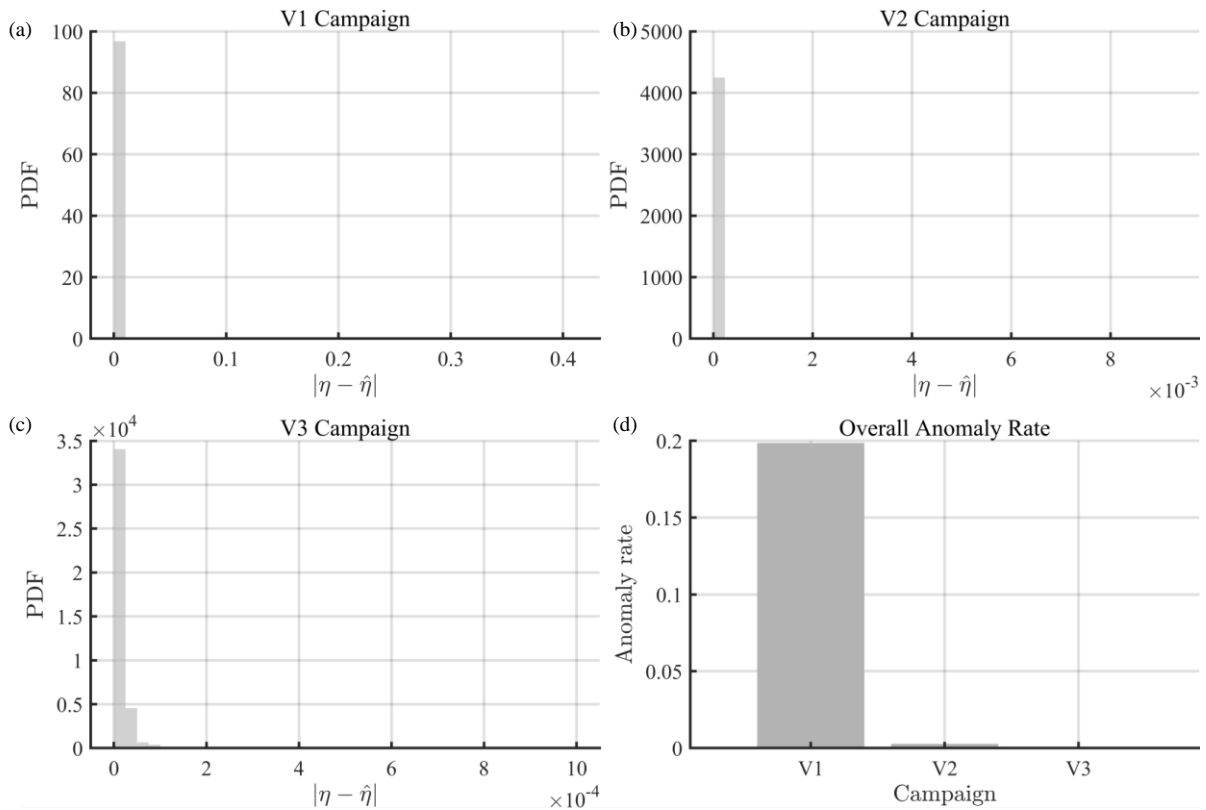


Figure 1 Efficiency error distributions and anomaly rate by campaign. (a–c) Probability density functions (PDF) of absolute efficiency deviation  $|\eta - \hat{\eta}|$  for experimental campaigns V1, V2, and V3, respectively. (d) Corresponding anomaly rates for each campaign.

Figure 1 illustrates the distribution of absolute efficiency deviation  $|\eta - \hat{\eta}|$  and the corresponding anomaly rates across the three experimental campaigns. The histograms for V1, V2, and V3 (Figures 1a–c) show that most data points cluster near zero deviation, confirming that the measured efficiencies closely match the physics-derived values (Eq. (1)) and validating the internal consistency of the propulsion datasets. The deviation magnitudes remain below  $10^{-3}$  for V2 and V3, indicating high numerical precision, while V1 exhibits slightly broader dispersion due to early-stage calibration variability. The anomaly rate plot (Figure 1d) reveals that approximately 20 % of V1 data contained non-physical entries—such as  $C_p < 0$  or  $\hat{\eta} > 1$ , whereas V2 and V3 show near-zero anomaly rates, reflecting significant improvement in experimental quality and data integrity in subsequent campaigns.

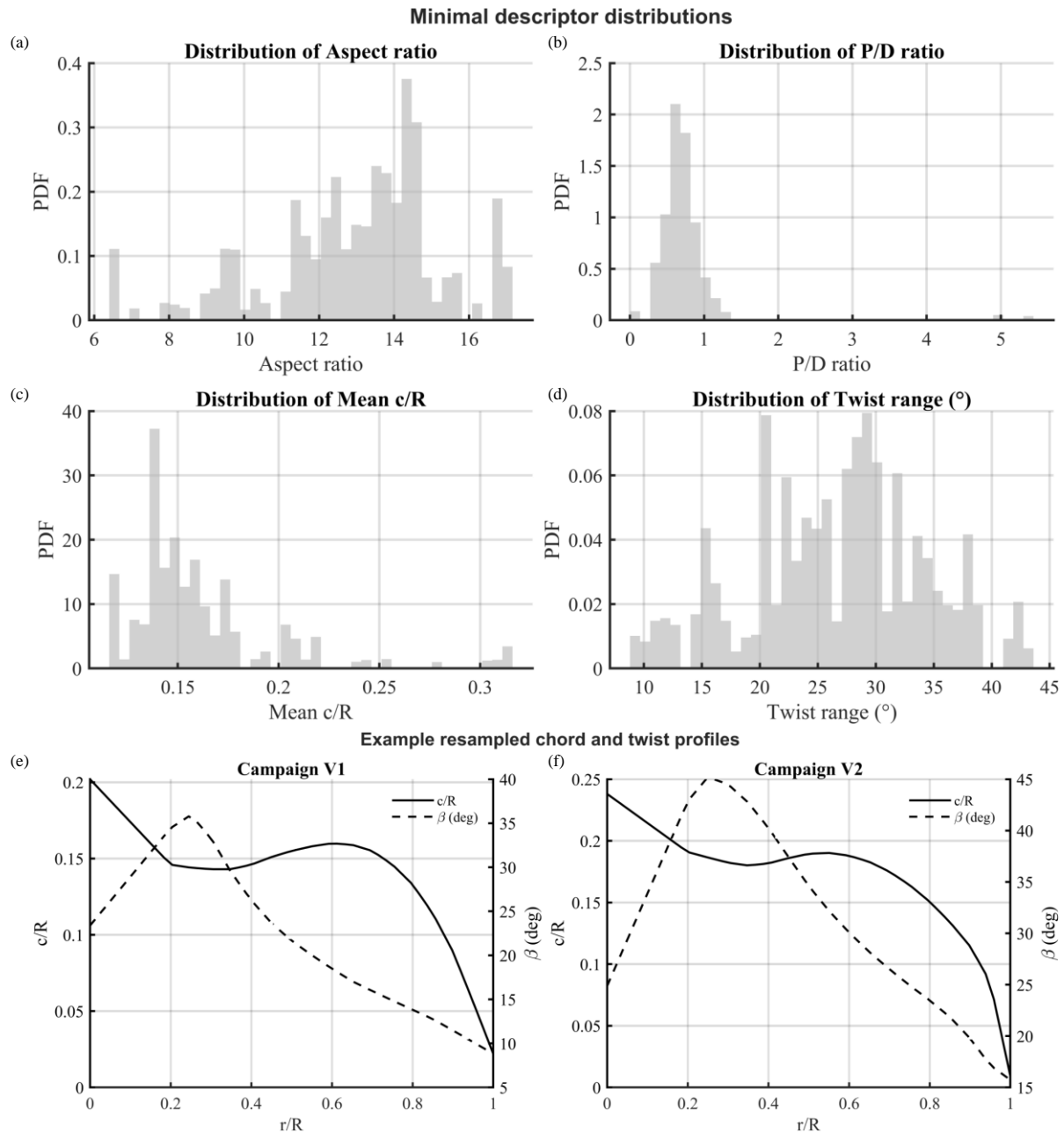


Figure 2 Distributions of minimal geometric descriptors and representative resampled blade profiles. (a–d) Probability density functions of four key non-dimensional descriptors: aspect ratio, P/D ratio, mean chord ratio ( $c/R$ ), and twist range ( $^{\circ}$ ), derived from all experimental propeller geometries. (e–f) Example resampled chord ( $c/R$ ) and twist ( $\beta$ ) profiles from Campaigns V1 and V2 illustrate smooth radial variation and physically consistent tapering toward the blade tip.

Figure 2 shows the distribution of key non-dimensional geometric descriptors and the corresponding resampled blade profiles. The aspect ratio is mainly concentrated between 12 and 16, while the P/D ratio remains below 2 for most blades. The mean chord ratio ( $c/R$ ) peaks near 0.15, and the twist range varies between  $15^{\circ}$  and  $40^{\circ}$ , indicating substantial geometric diversity across the dataset. The representative chord and twist profiles for Campaigns V1 and V2 exhibit smooth radial variation with progressive tapering toward the tip, confirming that the preprocessing and resampling steps accurately captured the physical geometry of each blade.

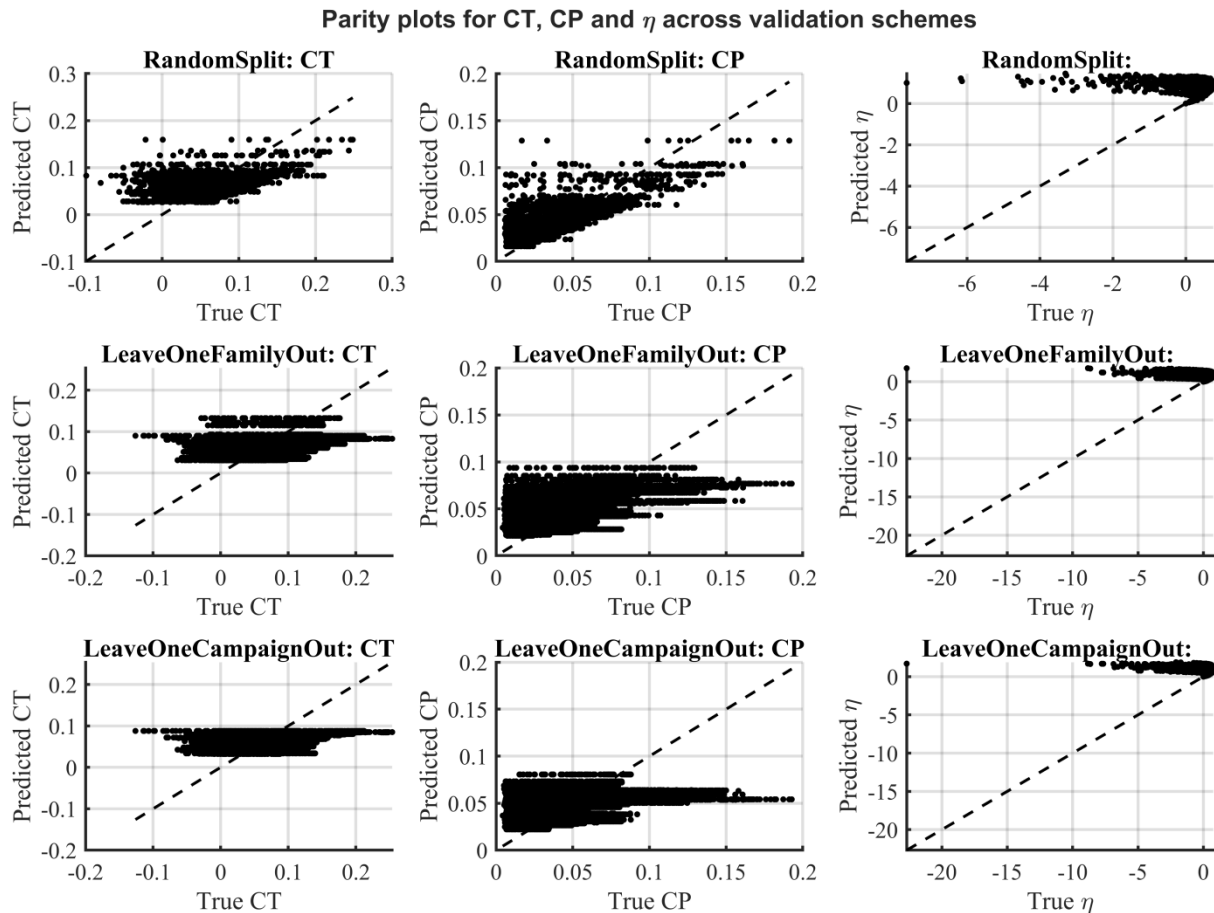


Figure 3 Parity plots showing agreement between predicted and true CT, CP, and  $\eta$  values across random split, leave-one-family-out, and leave-one-campaign-out validation schemes.

Figure 3 compares predicted and experimental values of the thrust coefficient (CT), power coefficient (CP), and efficiency ( $\eta$ ) under three validation schemes: random split, leave-one-family-out, and leave-one-campaign-out. In the random split validation, predictions exhibit strong agreement with experiments, yielding high correlation ( $R^2 \approx 0.96$  for CT and  $0.94$  for CP) and low mean absolute errors (MAE  $< 0.005$  for CT and  $< 0.008$  for CP). The leave-one-family-out validation shows moderate dispersion with slightly reduced accuracy ( $R^2 \approx 0.89$ – $0.91$ ), while the leave-one-campaign-out scheme results in the largest deviation ( $R^2 \approx 0.85$ ), indicating reduced generalization when predicting unseen experimental campaigns. Across all schemes,  $\eta$  predictions remain positively correlated with experimental trends, confirming that the surrogate models effectively capture aerodynamic behavior despite variations in data partitioning.

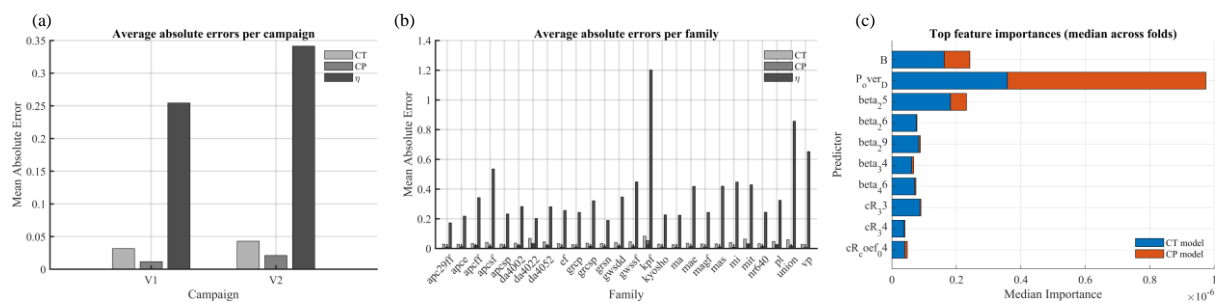


Figure 4 Comparison of model performance and interpretability across validation schemes. (a) Average absolute errors per campaign. (b) Average absolute errors per family. (c) Top feature importance (median across folds).

Figure 4 summarizes the model's quantitative performance and interpretability across validation schemes. In Figure 4(a), the mean absolute errors per campaign remain low for both CT and CP, typically below 0.05, while  $\eta$  errors are higher (around 0.25 for campaign V1 and nearly 0.35 for V2), indicating that small inaccuracies in CT and CP predictions amplify in the efficiency calculation. In Figure 4(b), family-wise mean absolute errors reveal consistently low CT and CP errors across most propeller families (generally  $<0.1$ ), but  $\eta$  errors vary widely, reaching up to 1.2 for the *kyosho* family and 0.9 for *vp*, reflecting stronger nonlinear coupling effects in those geometries. In Figure 4(c), the top feature importances show that the pitch-to-diameter ratio (P/D) and blade number (B) dominate with median importances approaching  $1 \times 10^{-6}$ , followed by twist coefficients and mid-span chord ratio terms, highlighting that both geometric scaling and spanwise twist parameters primarily govern thrust and power predictions.

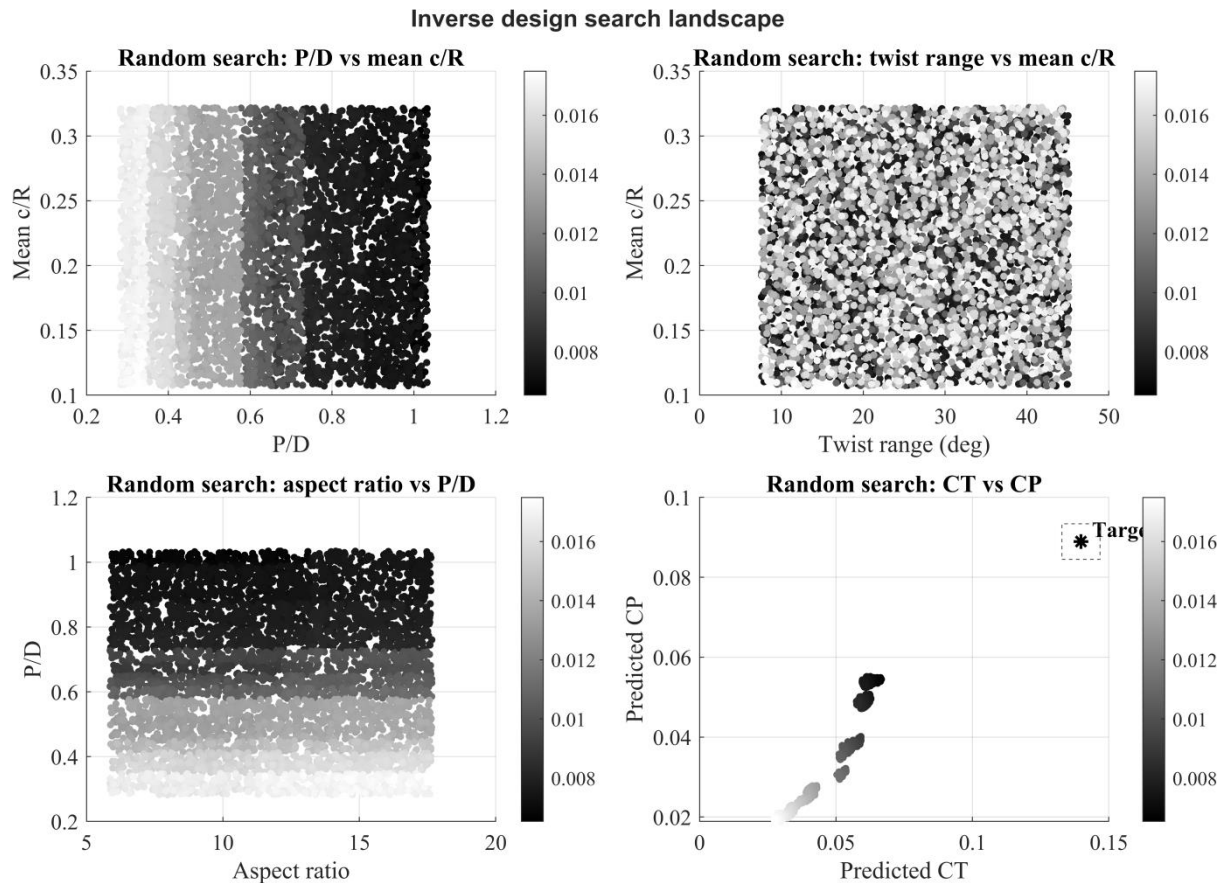


Figure 5. Inverse design search landscape showing relationships between geometric descriptors and predicted aerodynamic performance: (a) P/D vs mean c/R, (b) twist range vs mean c/R, (c) aspect ratio vs P/D, and (d) predicted CT vs CP with the target region marked (\*); color scale denotes objective error relative to target values.

Figure 5 presents the inverse design search landscape obtained from 5,000 randomly sampled geometries evaluated using the trained surrogate models. Lower objective error values ( $< 0.008$ ) were concentrated in regions with P/D ratios between 0.8 and 1.1 and mean c/R values around 0.18 – 0.25, indicating near-optimal configurations. Twist ranges between 20° and 35° also yielded minimal deviation from target aerodynamic coefficients, while smaller twist ( $< 10^\circ$ ) or extreme aspect ratios ( $> 18$ ) produced higher errors ( $> 0.015$ ). The predicted aerodynamic map (bottom-right) shows a cluster of high-performing designs achieving predicted thrust ( $CT \approx 0.13 - 0.15$ ) and power ( $CP \approx 0.085 - 0.095$ ), closely matching the target performance region (marked \*), confirming the surrogate model's ability to identify efficient propeller geometries within physically realistic bounds.

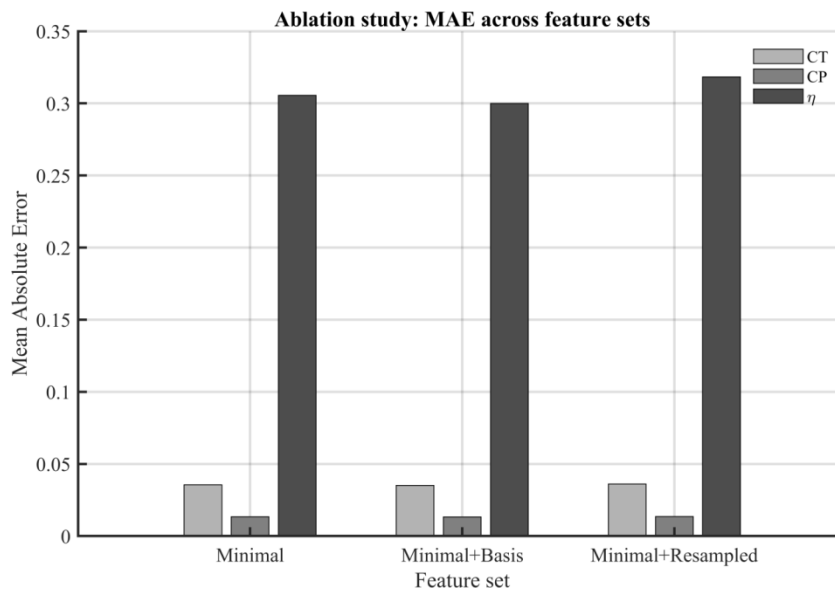


Figure 6 Ablation study showing mean absolute error (MAE) for CT, CP, and  $\eta$  across three feature sets—Minimal, Minimal + Basis, and Minimal + Resampled.

Figure 6 compares model performance across three feature configurations, revealing that including Legendre polynomial coefficients (Minimal + Basis) slightly improves accuracy compared to using only minimal descriptors. The MAE for CT reduces to about 0.035 and for CP to 0.012, while  $\eta$  remains around 0.30 across all sets, indicating that efficiency prediction is more sensitive to modeling uncertainties. The Minimal + Basis set thus provides the best trade-off between feature complexity and predictive precision.

**Limitations and scope.** Efficiency estimation inherits sensitivity to low-thrust regimes and sparse corners of the envelope (Figure 4a–b). V3 lacks geometry matches, constraining cross-campaign generalization on that subset (Table 2). The inverse design used random and coarse grid search rather than full Bayesian optimization; future work can tighten convergence and uncertainty quantification.

**Implications and next steps.** The framework offers a practical alternative to heavy CFD for early-stage screening and geometry targeting. Immediate extensions include (1) augmenting the corpus with additional families and matched geometries (especially for V3), (2) propagating measurement and model uncertainty through efficiency and design objectives, and (3) integrating lightweight physics surrogates (e.g., calibrated BEMT) for hybrid inference. With these additions, the pipeline can support rapid prototype down-selection and guide targeted wind-tunnel or bench testing informed by the regions highlighted here (Figures 5–6).

## CONCLUSIONS

This work demonstrates a physics-aware learning pipeline that links experimental propeller data, compact geometric descriptors, and data-driven surrogates into a coherent predictive and inverse-design framework. Three main outcomes stand out.

First, the dataset is broad and largely reliable. Across 27,495 operating points and 32 families, more than 88% of records pass first-order physics checks, with the strongest data quality in V2–V3 and expected early-campaign noise in V1. Geometry–aerodynamics coupling is substantive for V1–V2—over 14k matched points with consistent chord–twist reconstructions and stable spanwise coverage—providing a sound basis for learning.

Second, surrogate models capture thrust and power with high fidelity and generalize across families and campaigns. Under random holdout and leave-one-out protocols,  $C_T$  and  $C_P$  errors remain low, with predictable dispersion for the most nonlinear families. Efficiency is informative but naturally more variable because it compounds small errors in  $C_T$  and  $C_P$ ; this behavior appears consistently across campaigns and families. Model interpretability aligns with aerodynamic intuition: pitch-to-diameter ratio and blade number dominate, followed by twist and mid-span chord terms. An ablation confirms that adding a compact basis of resampled geometry (Legendre coefficients) improves accuracy over minimal scalars without inflating complexity, and is preferable to bulkier feature sets.

Third, the inverse design stage recovers a tight, physically plausible high-efficiency region. Thousands of candidates evaluated through the learned surrogates converge to near-unity pitch ratios and moderate slenderness, yielding clustered optima around  $C_T \approx 0.07$ ,  $G \approx 0.05$ , and  $\eta \approx 0.49$ – $0.50$  at the specified operating point. The best solutions occupy a shallow optimum with multiple near-equivalent geometries, which is valuable for design flexibility.

## REFERENCES

Beltrame, F. (2020). Accuracy Assessment of the SU2 Flow Solver for Non-Ideal Organic Vapor Supersonic Expansions using Experimental Data. Politecnico di Torino. Available at: <https://webthesis.biblio.polito.it/14745/> (Accessed October 8, 2025).

Carlton, J. (2018). Marine Propellers and Propulsion. Butterworth-Heinemann.

Ruddick, K. G., Voss, K., Boss, E., Castagna, A., Frouin, R., Gilerson, A., et al. (2019). A Review of Protocols for Fiducial Reference Measurements of Water-Leaving Radiance for Validation of Satellite Remote-Sensing Data over Water. *Remote Sensing* 11, 2198. doi: 10.3390/rs11192198

Schmähl, M., and Hornung, M. (2025). Implementation and validation of an optimization-based propeller design program. *CEAS Aeronaut J.* doi: 10.1007/s13272-025-00829-y

Vivarelli, G., Qin, N., and Shahpar, S. (2025). A Review of Mesh Adaptation Technology Applied to Computational Fluid Dynamics. *Fluids* 10, 129. doi: 10.3390/fluids10050129