

Campaign-resolved evidence for adaptive critical-heat-flux surrogates under experimental shift

Satya Prakash Saraswat

`satya.saraswat@unitus.it`

University of Tuscia

Article

Keywords: critical heat flux, campaign-held-out validation, experimental provenance, physics-anchored machine learning, negative transfer, directional data valuation, conformal prediction, sparse Bayesian assimilation, adaptive surrogate

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10789551/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Campaign-resolved evidence for adaptive critical-heat-flux surrogates under experimental shift

Satya Prakash Saraswat^{1,2, *}

¹ School of Mechanical Engineering, Department of Economics, Engineering, Society and Business Organization (DEIM), University of Tuscia, 01100 Viterbo, Italy

²Department of Physics, Division of Nuclear Science and Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

* Correspondence: satya.saraswat@unitus.it | ORCID: 0000-0003-2007-8158

Research Article for the Collection 'Artificial Intelligence for Advancing Thermal Science and Engineering'

Abstract

Critical heat flux (CHF) is a limiting heat-transfer condition in water-cooled systems. Prediction has progressed from empirical correlations and lookup tables to mechanistic models, physics-informed machine learning, and uncertainty-aware hybrids. These advances improve predictive capability, but a separate question remains: whether knowledge learned from experiments remains reliable when an entire experimental campaign, with its own test section, instrumentation, surface history, and operating trajectory, is absent from model development.

This study treats the campaign as the common unit of generalization, uncertainty calibration, evidence valuation, and sparse adaptation. Using 24,579 water-CHF measurements from 60 uniformly heated vertical round-tube campaigns, physics-anchored and direct machine-learning models are tested under exhaustive whole-campaign separation. The same campaign structure is then examined through conformal uncertainty, directed source-to-target transfer, coalition data value, and regularized Bayesian updating from a few revealed measurements. An independent hierarchical response-field analysis tests whether campaign-level residual structure persists under a different estimator, while a one-dimensional energy balance is used only as a conservation check.

The analysis shows that whole-campaign prediction is harder than record-level prediction, and historical sampling concentration explains only part of that difference. Cross-campaign evidence is not exchangeable: usefulness depends on thermohydraulic support, can reverse with transfer direction, and yields coalition rankings that depend on the transport rule. Record-level uncertainty calibration does not provide uniform protection across campaigns. By contrast, a small number of new measurements can recover part of the unresolved campaign-specific state, although the acquisition policy does not statistically outperform simple random sampling.

The work, therefore, reframes CHF machine learning as an experimental evidence transport problem. The resulting framework is an offline, campaign-initialized adaptive surrogate for deciding when historical evidence is transferable, when uncertainty should be widened, and when limited new measurements can support prediction updates; it does not claim to identify mechanisms, achieve multiphase simulation fidelity, or attain operational digital twin status.

Highlights

1. Campaign provenance remains relevant to CHF generalization.
2. Cross-campaign evidence value is directional, not symmetric.
3. Thermohydraulic support shapes helpful and harmful transfer.
4. Campaign-level uncertainty requires campaign-aware calibration.
5. Few measurements recover campaign residual state; complex querying adds no proven gain.

Keywords: critical heat flux; campaign-held-out validation; experimental provenance; physics-anchored machine learning; negative transfer; directional data valuation; conformal prediction; sparse Bayesian assimilation; adaptive surrogate

Acknowledgments

*This work was carried out within the **PROTEZIONE project “Machine Learning-Driven Safety Enhancement for Critical Heat Flux Prediction in Nuclear Reactors,”** led by the author as Principal Investigator. PROTEZIONE is funded by the European Union – NextGenerationEU under the Italian National Recovery and Resilience Plan (PNRR), the Young Researchers 2024 Program, and the Marie Skłodowska-Curie Actions line, and is implemented by the Italian Ministry of University and Research. Project code: MSCA2024_0000004; CUP: J87G25000150007. The funder had no role in the study design, data analysis and interpretation, preparation of the manuscript, or the decision to submit the work for publication.*

1. Introduction

1.1 Physical problem and evidentiary unit

Critical heat flux (CHF) marks a limiting condition in boiling heat transfer at which heat-removal capability can deteriorate sharply, and the heated wall can experience a rapid temperature rise. It is therefore a central thermal-design and safety quantity in water-cooled reactor systems [1]. Experimental and modeling reviews emphasize that CHF does not arise from one universal microscopic pathway; under subcooled flow-boiling conditions, much mechanistic work focuses on departure from nucleate boiling (DNB) and the near-wall processes associated with it [2]. Mechanistic DNB models have been formulated from local bulk-flow conditions [3], while other work addresses CHF behavior at low qualities [4]. Hejzlar and Todreas further examined how subcooled and low-quality CHF correlations affect predicted CHF margin [5]. Because the archive used here does not contain direct measurements of the microscopic triggering process, this study does not assign any statistical response pattern to a specific DNB or dryout mechanism.

Engineering prediction has developed in parallel through database tools and correlations. The 2006 Groeneveld lookup table is a normalized CHF database for a vertical water-cooled tube, derived from a large experimental database [6]; the data used to generate that table were later documented separately by the U.S. Nuclear Regulatory Commission [7]. Classical round-tube prediction methods include Bowring [8] and Biasi [9], as well as the Katto-Ohno [10] and Shah [11] correlations. Tong addressed DNB prediction for axially nonuniform heat flux [12], whereas Zhang et al. developed a water-CHF correlation for mini-channels [13]. Broader thermohydraulic compilations summarize engineering relationships used in advanced water-cooled reactor analysis [14]. Comparative assessments of subcooled CHF correlations likewise show that predictive behavior depends on operating range and model formulation [15,16]. A saturated-flow-boiling correlation for mini/microchannels provides a separate example of geometry-specific closure development outside the present round-tube CHF target [17]. Surface modification is also known to alter flow-boiling behavior, which makes surface history a plausible but incompletely recorded campaign attribute [18].

These physics and data-modeling issues make experimental provenance relevant to validation. Rows from one campaign share a test section, instrumentation chain, heater, and surface history, inlet-conditioning system, and operating trajectory, so the campaign is a natural grouping variable for the transfer question studied here. Cross-validation is widely used for generalization assessment and model selection [19,20]. When observations have temporal, spatial, hierarchical, or other dependence, the resampling strategy should respect that structure [21]. Model-selection studies also show that estimated performance can become optimistic when model selection and final evaluation are not kept separate [22,23]. Accordingly, this study treats the campaign, not the individual database row, as the unit that must remain unseen during final evaluation.

1.2 State of the art: from CHF correlations to uncertainty-aware machine learning

CHF machine learning has progressed from standalone regression to physics-assisted and uncertainty-aware models. Zhao et al. proposed a physics-informed machine-learning-aided framework for CHF prediction [24]. Khalid et al. compared standalone and hybrid machine-learning models for vertical tubes [25], and Helmryd Grosfilley et al. investigated several regression techniques over a large CHF parameter space [26]. Furlong et al. combined Bowring or Biasi reference models with deep neural-network ensembles, Bayesian neural networks, and deep Gaussian processes and evaluated both prediction and uncertainty [27]. Mao and Jin focused specifically on uncertainty quantification for physics-informed CHF models [28]. Alsafadi et al. studied conditional variational autoencoders and deep neural networks, including uncertainty quantification and domain generalization tests [29]. Zhang et al. extended hybrid CHF prediction through stacked learning [30]. Cazzola et al. examined coverage-oriented uncertainty quantification for learning complex CHF response regimes [31]. Dunbar et al. showed that tube-based CHF prediction methods can perform poorly when transferred to internally heated annuli, underscoring that geometry transfer is a separate problem [32].

These CHF studies sit within a broader scientific-machine-learning literature in which physical knowledge and data-driven learning are combined rather than treated as mutually exclusive [33]. Physics-informed formulations include extensions such as fPINNs for fractional partial differential equations [34]. Reviews in fluid mechanics describe opportunities and limitations for machine learning in data-driven modeling [35], while turbulence-modeling literature discusses how data can augment or correct traditional closure models

[36]. Nuclear-engineering benchmark programs likewise illustrate the value of common problem definitions for uncertainty comparison; the OECD/NEA LWR-UAM Phase-I benchmark is one example [37]. A recent nuclear-engineering review emphasizes that machine-learning prediction uncertainty can arise from data noise, limited coverage, extrapolation, model approximation, and stochastic training [38].

Validation design remains central to the present question. Ordinary record-level random splitting and whole-campaign testing answer different questions when rows from the same experiment share apparatus and operating history. Group-aware validation is therefore used here not because random splitting is intrinsically invalid, but because the deployment target is a campaign absent from model development. The model-selection firewall is equally important: preprocessing, learner choice, ensemble weighting, and interval calibration are all determined without consulting the target campaign. Physics anchoring does not remove this requirement, because a residual learner can still absorb a campaign-specific discrepancy if that campaign is represented during development.

Uncertainty has the same unit-of-evidence issue. Conformal prediction provides finite-sample marginal predictive guarantees under its stated exchangeability assumptions [39,40], while jackknife+ provides another predictive-inference construction [41]. Work on conformal prediction under covariate shift shows that distribution change requires additional weighting or assumptions [42]. Classical tolerance-limit theory likewise distinguishes a population-level statement from a descriptive sample statistic [43]. Deep ensembles provide a practical, scalable approach to predictive uncertainty estimation [44]. For this study, the relevant question is therefore not only whether intervals cover pooled records, but whether they protect a typical held-out campaign when campaigns contribute very unequal numbers of measurements.

Historical data value is a different problem from feature attribution. SHAP explains how features contribute to an individual prediction [45], and dependent-feature work shows that such explanations depend on the assumed feature distribution [46]. Data Shapley instead values training observations through their marginal contribution to model utility [47], with subsequent work addressing more efficient Shapley-based data valuation [48]. Transfer-learning literature formalizes knowledge reuse across domains [49], while negative-transfer studies explicitly consider cases in which source information harms target performance [50]. Importance-weighted cross-validation addresses covariate shift under stated assumptions [51], and influence functions provide another way to estimate how training data affect predictions [52]. These ideas motivate the campaign-level question tested here: Can campaign A help predict campaign B, whereas the reverse is harmful?

That question connects directly to sparse adaptation. If pressure, mass flux, inlet state, and geometry do not fully capture the residual state associated with a new experiment, a few newly observed CHF values may contain information unavailable from historical inputs alone. The scientific test is deliberately split into two parts: whether sparse target measurements improve prediction for the remainder of the campaign, and whether a sophisticated acquisition rule selects those measurements better than simple alternatives. Improvement after an update is evidence of adaptation; it is not, by itself, evidence that an active-learning policy is optimal.

Digital-twin terminology provides an equally important boundary. Wright and Davidson distinguish a digital twin from an ordinary model by its relationship to data from a physical system and its ability to be dynamically updated [53]. Rasheed et al. describe a digital twin as a virtual representation of a physical asset, enabled by data and simulators, for real-time prediction, monitoring, control, optimization, and decision support [54]. Thelen et al. review physical-to-virtual and virtual-to-physical data flows as core technologies that enable twinning [55]. The present study has no live asset, continuous sensor connection, or synchronized physical-to-virtual state. It replays completed campaigns offline and updates only a statistical CHF residual after selected measurements are revealed. We therefore describe it as an offline adaptive campaign-initialized surrogate; TWIN-VALUE is retained only as the archived workflow and acquisition-policy label. Table 1 summarizes the main areas of the literature that frame the present campaign-resolved CHF analysis and clarifies their roles in the study.

Table 1 | Literature context for the campaign-resolved CHF analysis.

Literature area	Relevance to the present study
CHF correlations and lookup tables [6,8]	Establish the conventional engineering basis for CHF prediction and provide the reference context for the physics-anchored surrogate.
Mechanistic and regime-specific CHF models [2,3]	Provide the physical background of CHF and DNB, and define the boundary between predictive residual structure and mechanism identification.
CHF machine learning and hybrid models [24,27]	Establish data-driven and physics-assisted CHF prediction as the broader methodological setting of the present work.
Grouped validation and model selection [21,23]	Provide the statistical basis for treating dependent experimental groups separately during model development and evaluation.
Uncertainty quantification [39,42]	Provides the foundation for predictive calibration and for examining uncertainty under changes in experimental support.
Data valuation and transfer [47,50]	Provides the conceptual basis for distinguishing source usefulness, directional transfer, and negative transfer between experiments.
Model updating and digital-twin concepts [53,55]	Provide the broader context for data-assisted model updating while clarifying the distinction between offline adaptation and operational digital twinning.

1.3 Research gap and falsifiable hypothesis

The gap addressed here is therefore not a lack of accurate CHF correlations, mechanistic ideas, machine-learning models, or uncertainty tools. It is the absence, for this 60-campaign archive, of one common evidentiary unit across the full chain of inference. A record can be held out while its campaign remains represented; an interval can be calibrated over records while a facility is poorly covered; a source can be useful on average yet harmful to a particular target; and a sparse update can work even when the rule used to select measurements adds no benefit. This study asks what remains true when the campaign is used consistently for generalization, uncertainty, evidence transfer, and adaptation.

The central hypothesis is predictive rather than causal: after the recorded thermohydraulic and geometric coordinates are supplied, experimental provenance retains a reproducible residual component that changes how evidence transfers, how uncertainty is distributed across campaigns, and how a new experiment responds to a few revealed measurements. The hypothesis is deliberately falsifiable. It would be weakened if record-weighted and campaign-weighted conclusions became equivalent after a weight-only concentration audit; if reciprocal campaign transfer were effectively symmetric; if directed transfer signs were unstable across reasonable transport games; if an independent hierarchical estimator found no reproducible campaign-level structure; if campaign-aware calibration did not improve facility-level coverage; or if regularized few-shot updating failed to reduce error.

Campaign identity is not treated as a thermophysical mechanism. It is a provenance label that may summarize unrecorded differences in surface preparation and aging, heater construction, geometric tolerance, instrumentation and calibration, inlet development, operating protocol, and the mixture of response patterns sampled by an experiment. Terms such as negative transfer, nuisance-explained edge, residual-conflict candidate, and response pattern are therefore only predictive labels. They do not diagnose DNB, dryout, surface bias, or measurement error without independent experimental evidence.

1.4 Objectives and contribution

Objective 1: Build a leakage-audited prediction framework in which every campaign is tested exactly once and all preprocessing, model selection, ensemble weighting, and interval calibration are determined from disjoint training and validation campaigns.

Objective 2: Quantify the directed value of every source campaign for every target campaign, test reciprocal symmetry, and determine how thermohydraulic distance, relative support, campaign size, and the chosen transport game affect that value.

Objective 3: Compare pooled-record and campaign-aware conformal calibration using record-weighted, equal-campaign, lower-tail, and campaign-specific coverage measures, without treating retrospective coverage as a universal guarantee for a new facility.

Objective 4: Replay the initialization of unseen campaigns with 1, 3, 5, and 10 revealed measurements, separating the benefit of regularized Bayesian residual updating from the performance of TWIN-VALUE, maximin, uncertainty-driven, and repeated-random acquisition policies.

Objective 5: Test whether the campaign signal survives an independent hierarchical TRACE analysis, couple predicted CHF only to a verified one-dimensional energy balance, and freeze a prospective 30-condition falsification matrix without implying mechanism identification, CFD fidelity, live synchronization, or operational digital-twin validation.

The contribution is therefore methodological and evidentiary rather than a new CHF correlation. The study makes the experimental campaign the common unit of evidence and asks whether conclusions survive that choice across prediction, calibration, source value, and sparse updating. It also separates claims that are often blended: prediction accuracy from transferability, data value from causality, adaptation from acquisition-policy superiority, response topology from mechanism identification, and conservation checking from multiphase simulation. The resulting evidence chain is designed so that important claims can fail, and the failed acquisition-superiority gate is retained as part of the scientific result.

1.5 Scope and claim boundaries

The scope is water CHF in uniformly heated vertical round tubes, as shown in the locked archive. The results do not establish external validity for other fluids, annuli, rod bundles, compact channels, nonuniform axial heating, transient operation, or facilities outside the represented support. The study does not identify a boiling-crisis mechanism, infer causal effects on facilities, prescribe a licensing margin, or establish an operational digital twin. Campaign-aware conformal coverage and sparse updating are retrospective results on archived experiments. The one-dimensional thermal calculation verifies only the stated energy balance, and the 30-condition matrix is a frozen proposal for future falsification rather than completed prospective validation.

2. Methods

2.1 Locked data, physical conditions, and statistical unit

The analysis uses the publicly documented water round-tube CHF compilation associated with the 2006 Groeneveld look-up-table evidence base [6,7]. The screened comma-separated file is fixed and contains 24,579 records assigned to 60 reference-identifier (RID) campaigns. No row is synthesized, imputed, or duplicated. Positive pressure, mass flux, diameter, heated length, and CHF are enforced; all required variables are complete.

RID is the unit of statistical transport. It is used only to allocate folds, balance training weights, aggregate metrics, and report campaign-specific results. RID is never a predictor. Likewise, measured CHF and any outlet quantity reconstructed from measured CHF are prohibited from the predictor interface. Saturation properties are evaluated as a function of pressure using a frozen IAPWS-IF97 cache [56], so repeated executions use identical thermodynamic states.

Campaign concentration is quantified because it changes the scientific meaning of record-weighted performance. The largest campaign contains 5,252 records; the five largest contain 54.0% of the archive; and the 30 smallest contain 5.9%, giving a record-concentration Gini coefficient of 0.719. This imbalance reflects the database's historical construction, not the physical importance of a campaign. It therefore motivates equal-campaign estimands and a campaign-balanced training candidate, but is not interpreted as a thermophysical finding.

Dataset-balance audit. Three safeguards are used against dominance by large historical campaigns: every outer split is performed by RID; headline performance is reported both per record and with equal campaign weight; and `ET_RESIDUAL_BAL` gives every training campaign equal total weight. In addition, a weight-only sensitivity is applied to the frozen cross-fitted champion predictions: campaigns are divided into terciles of equal count, and record-weighted versus equal-campaign MAPE is compared within each tercile. Then, every campaign contribution is capped at 500 records, and the weighted MAPE is recomputed without retraining. This isolates the part of the record-to-campaign discrepancy attributable to reporting weights from the part that remains as unequal campaign difficulty. It is not a counterfactual retraining experiment and is not interpreted as a physical decomposition.

2.1.1 Concentration-versus-difficulty estimand audit

Let MAPE_{rw} denote the equal-record mean and $\text{MAPE}_{\text{macro}}$ the equal-campaign mean of the same frozen cross-fitted predictions. Within each campaign-size tercile, their ratio measures how much unequal within-tercile record counts alter the estimand while holding broad size class fixed. The 500-record cap supplies a

second weight-only perturbation. The fraction of the observed champion gap recovered by reweighting is defined as $(\text{MAPE_cap500} - \text{MAPE_rw})/(\text{MAPE_macro} - \text{MAPE_rw})$; the complement is reported conservatively as residual campaign-difficulty heterogeneity. Because no model is retrained, neither component is described as a causal or thermophysical fraction. The locked database variables, physical ranges, and their permitted analytical roles are summarized in Table 2.

Table 2 | Locked database, variable definitions, physical envelope, and permitted analytical role.

Variable or quantity	Symbol and unit	Range or value	Use and constraint
Records	N	24,579	Prediction rows; never treated as independent campaigns
Campaigns	J	60 RIDs	Split, bootstrap, and macro-aggregation unit
Outer blocks	B	12	Load-balanced; each RID tested exactly once
Pressure	P, MPa	0.10–20.0	Predictor: saturation properties depend only on P
Mass flux	G, kg m ⁻² s ⁻¹	8.2–7,964	Predictor
Inner diameter	D, mm	2.0–16.0	Predictor
Heated length	L, m	0.05–20.0	Predictor
Heated-length ratio	L/D	25–2,484	Dimensionless geometry predictor
Inlet enthalpy deficit	$\Delta h_{\text{sub, in}}$, MJ kg ⁻¹	–1.211 to 1.644	Predictor retained as archived
Measured CHF	q_{CHF} , MW m ⁻²	0.050–16.339	Positive target; forbidden from selection inputs
Campaign size	n_j	1–5,252	Weighting and audit only; not a predictor
The largest five campaigns	$\Sigma n(5)/N$	54.0%	Evidence-concentration diagnostic
Record-concentration Gini	Gini(n_j)	0.719	Motivates campaign-level estimands
Bowring raw outputs at or below the floor	count	29 records	Floored only for finite log residual; sensitivity reported

2.2 Thermophysical preprocessing and feature construction

Let $P_{\text{crit}} = 22.064$ MPa, $G_0 = 1000$ kg m⁻² s⁻¹, $D_0 = 0.01$ m, $L_0 = 1$ m, $h_0 = 10^6$ J kg⁻¹, and $q_0 = 1$ W m⁻². From IAPWS-IF97, the liquid density ρ_f , vapor density ρ_g , latent heat h_{fg} , liquid viscosity μ_f , surface tension σ , and saturation temperature T_{sat} are cached for each observed pressure. The reference scales make logarithmic inputs dimensionless while preserving the implemented feature scaling. The Bowring anchor is evaluated with SI inputs. Its principal form is

$$q_B = \frac{A + \frac{DG \Delta h_{\text{sub, in}}}{4}}{C + L} \quad (1)$$

$$A = \frac{2.317 \left(\frac{h_{fg} DG}{4} \right) f_1}{1 + 0.0143 f_2 G D^{1/2}} \quad C = \frac{0.077 f_3 DG}{1 + 0.347 f_4 \left(\frac{G}{1356} \right)^n}$$

where $n = 2 - 0.5 p_r$ and $p_r = P/(6.895 \text{ MPa}) = 0.145(P/\text{MPa})$; the piecewise pressure functions f_1 – f_4 follow Bowring [8] exactly and are preserved in the Supplementary Software. A raw Bowring value ≤ 1 kW m⁻² is replaced by 1 kW m⁻² only to keep the logarithm finite. All 29 affected rows are exported, and the exclusion sensitivity is reported.

For record i , the model vector contains ten pre-CHF quantities chosen before outer testing:

$$\mathbf{u}_i = \left[\ln(P/P_{\text{crit}}), \ln(G/G_0), \ln(D/D_0), \ln(L/L_0), \ln(L/D), \Delta h_{\text{sub,in}}/h_{fg}, \Delta h_{\text{sub,in}}/h_0, \ln(q_B/q_0), \ln(GD/\mu_f), \ln(G^2 D/(\rho_f \sigma)) \right]^T \quad (2)$$

Within every outer block, each feature is standardized with the mean and standard deviation of the outer-training rows only. Validation and test rows are transformed with those frozen statistics. The resulting standardized feature vector is denoted $\tilde{\mathbf{u}}_i$. Thus, no target-campaign distributional information enters preprocessing.

2.3 Campaign-resolved outer testing and information firewall

Campaigns are greedily assigned to twelve load-balanced outer-test blocks using record counts and seed 20260814. Within block B, the remaining campaigns are shuffled deterministically, and approximately 15% are assigned to validation, with a minimum of three validation campaigns. Training, validation, and test RID sets are mutually exclusive. Across blocks, the concatenated test sets contain every RID exactly once.

The outer training campaigns determine the preprocessing and learner parameters. Validation campaigns determine the candidate champion, the convex ensemble weights, and the conformal half-widths. Test campaigns are used only after these objects are frozen. Cross-fitted test predictions are concatenated after all 12 blocks have finished. The information firewall, therefore, applies to records, campaign labels, model selection, uncertainty calibration, and acquisition priors.

2.4 Physics anchor, learning targets, and candidate predictors

Each candidate model m learns a logarithmic target relative to a model-specific reference flux and reconstructs CHF through the inverse log transform:

$$\eta_i^{(m)} \equiv \ln \left(\frac{q_{\text{CHF},i}}{q_{\text{ref},i}^{(m)}} \right), \quad \hat{\eta}_i^{(m)} \equiv h_{\theta}^{(m)}(\tilde{\mathbf{u}}_i) \Rightarrow \hat{q}_i^{(m)} = q_{\text{ref},i}^{(m)} \exp(\hat{\eta}_i^{(m)}) \quad (3)$$

$$q_{\text{ref},i}^{(m)} = \begin{cases} q_{B,i}, & m \in \mathcal{M}_{\text{res}}, \\ q_0, & m \in \mathcal{M}_{\text{dir}}. \end{cases}$$

The logarithmic representation preserves positivity and makes multiplicative errors additive. The physical anchor supplies a transparent engineering baseline but does not constrain the residual learner to a mechanistic CHF law. Four candidates are trained on identical outer-training campaigns: record-weighted Extra Trees residual, campaign-balanced Extra Trees residual, record-weighted histogram gradient boosting residual, and record-weighted direct Extra Trees. Campaign-balanced weights are $w_i = N_{\text{train}}/(J_{\text{train}} n_j)$, so every training campaign contributes equal total weight; this member is also the primary protection against the historical record imbalance quantified in Section 2.1.

2.5 Frozen hyperparameters, validation-only stacking, and selection

Hyperparameters are fixed in the analysis package and are not searched against test performance [57,58]. Extra Trees uses 320 trees, minimum leaf size 1, no bootstrap, and all processors [59,60]; the campaign-balanced member uses $\text{max_features} = 0.80$, and the other Extra Trees members use 1.00. Histogram gradient boosting uses absolute-error loss [59,61], 420 iterations, a learning rate of 0.055, 31 terminal leaves, a minimum leaf size of 12, and L2 regularization of 0.5. Model-specific seed offsets are combined with the outer-block seed. LightGBM [62], XGBoost [63], and Random Forests [64] are cited as related tree- and boosting-based families but are not included in the frozen candidate set evaluated here.

A non-negative convex combination is fitted in log space using validation campaigns only:

$$\hat{q}_{\text{ens},i} = q_0 \exp \left[\sum_{m \in \mathcal{M}} w_m \ln \left(\frac{\hat{q}_i^{(m)}}{q_0} \right) \right], \quad w_m \geq 0, \quad \sum_{m \in \mathcal{M}} w_m = 1 \quad (4)$$

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmin}} \frac{1}{|V|} \sum_{j \in V} \text{MAPE}_j(\mathbf{w})$$

Sequential least-squares programming minimizes campaign-balanced validation MAPE subject to the simplex constraint. If optimization fails, the best single validation candidate receives a unit weight. The candidate with the minimum validation campaign-macro MAPE among the four learners and the convex ensemble is the block champion. The frozen candidate models, prediction targets, weighting schemes, estimator specifications, and validation-only selection rule are summarized in Table 3.

Table 3 | Frozen candidate models, targets, weighting, and selection rule.

Candidate	Prediction target	Training weight	Frozen estimator specification	Role
ET_RESIDUAL_A	$\ln(q_{CHF}/q_B)$	Equal records	Extra Trees: 320 trees; min leaf 1; max features 1.00; no bootstrap	Residual candidate
ET_RESIDUAL_BAL	$\ln(q_{CHF}/q_B)$	Equal campaigns	Extra Trees: 320 trees; min leaf 1; max features 0.80; no bootstrap	Campaign-balanced residual
HGB_RESIDUAL	$\ln(q_{CHF}/q_B)$	Equal records	Absolute loss; 420 iter.; $\eta=0.055$; 31 leaves; min leaf 12; $L_2=0.5$	Smooth residual candidate
ET_DIRECT	$\ln(q_{CHF}/q_0)$	Equal records	Extra Trees: 320 trees; min leaf 1; max features 1.00; no bootstrap	Anchor-free comparator
CONVEX_ENSEMBLE	Convex log prediction	Equal validation campaigns	SLSQP; $w_m \geq 0$; $\sum_m w_m = 1$; max 300 iter.; $f_{tol}=10^{-12}$	Validation-only stacked candidate
VALIDATION_CHAMPION	Lowest validation macro MAPE	One vote per validation campaign	Minimum of four learners and an ensemble	Only model used for headline test predictions

2.6 Performance estimands, selection regret, and inference

For campaign j , define the record-level absolute percentage error and the two aggregation conventions as

$$e_{ij} = 100 \left| \frac{\hat{q}_{ij}}{q_{CHF,ij}} - 1 \right|$$

$$\text{MAPE}_{\text{rw}} = \frac{1}{N} \sum_{j=1}^J \sum_{i=1}^{n_j} e_{ij}, \quad \text{MAPE}_{\text{macro}} = \frac{1}{J} \sum_{j=1}^J \left(\frac{1}{n_j} \sum_{i=1}^{n_j} e_{ij} \right) \quad (5)$$

Campaign median, campaign p90, and worst-campaign MAPE describe the distribution across the 60 within-campaign values. Signed bias, RMSPE, multiplicative k95, and fractions within $\pm 10\%$ and $\pm 20\%$ are also retained. Paired uncertainty intervals resample campaigns rather than records, using 4,000 bootstrap replicates [65]. Selection regret is the test MAPE of the validation champion minus the minimum test MAPE among candidates. This retrospective test-best comparator is reported only to quantify the cost of honest selection and never feeds back into development.

2.7 Pooled and campaign-robust conformal calibration

For validation records, the nonconformity score is the absolute log-residual. With n_{cal} calibration records and nominal level $1-\alpha$, the finite-sample pooled half-width is the higher of the two empirical quantiles.

$$a_i = \left| \ln \left(\frac{q_{CHF,i}}{\hat{q}_i} \right) \right|$$

$$c_\alpha = \text{Quantile}_{\text{higher}} \left(\{a_i\}_{i=1}^{n_{\text{cal}}}, \min \left[1, \frac{\text{ceil}[(n_{\text{cal}} + 1)(1 - \alpha)]}{n_{\text{cal}}} \right] \right) \quad (6)$$

$$I_\alpha(\mathbf{x}) = [\hat{q}(\mathbf{x})e^{-c_\alpha}, \hat{q}(\mathbf{x})e^{c_\alpha}]$$

The campaign-robust rule first computes the same within-campaign quantile $c_{j,\alpha}$ for each validation campaign and then takes the higher empirical $(1-\alpha)$ quantile across those campaign values:

$$c_{CR,\alpha} = \text{Quantile}_{\text{higher}}(\{c_{j,\alpha}: j \in V\}, 1 - \alpha) \quad (7)$$

$$I_{CR,\alpha}(\mathbf{x}) = [\hat{q}(\mathbf{x})e^{-c_{CR,\alpha}}, \hat{q}(\mathbf{x})e^{c_{CR,\alpha}}]$$

Because each block contains only about eight validation campaigns, the second quantile is intentionally conservative and often equals the maximum. Both rules are evaluated at the 90% and 95% nominal levels using record-weighted coverage, mean campaign coverage, lower-decile campaign coverage, and multiplicative interval width. These are retrospective empirical coverage statements for held-out members of the archived campaign population. Neither rule guarantees a stated coverage level for a genuinely new facility whose unrecorded surface, metrology, or response distribution lies outside those represented in the archive.

2.8 Directed transport games and removal value

Transfer is evaluated in a separate five-dimensional robustly standardized space comprising log pressure, log mass flux, log diameter, log L/D, and normalized inlet subcooling. Each source campaign s supplies a local expert for the Bowring log residual at the target transport coordinate ξ . The 12 nearest source records receive Gaussian within-source weights; source experts are then combined by one of three between-source games: Gaussian support, inverse distance, or the ten nearest campaigns. The three games deliberately span different smoothness assumptions: kernel-weighted continuity, distance-dominated support, and local-neighbor stationarity, so that agreement of sign is less dependent on a single transport rule.

$$\hat{r}_s(\xi) = \frac{\sum_{i \in \mathcal{N}_{12}^{(s)}(\xi)} K\left(\frac{\|\xi - \xi_i\|}{\ell_s}\right) \ln\left(\frac{q_{CHF,i}}{q_{B,i}}\right)}{\sum_{i \in \mathcal{N}_{12}^{(s)}(\xi)} K\left(\frac{\|\xi - \xi_i\|}{\ell_s}\right)} \quad (8)$$

$$\hat{r}_s^{(g)}(\xi) = \frac{\sum_{s \in \mathcal{S}} \omega_s^{(g)}(\xi) \hat{r}_s(\xi)}{\sum_{s \in \mathcal{S}} \omega_s^{(g)}(\xi)}, \quad g \in \mathcal{G}$$

For target t , the directed removal marginal of source s is

$$V_{\text{rem}}^{(g)}(s \rightarrow t) = \text{MAPE}_t^{(g)}(S \setminus \{s\}) - \text{MAPE}_t^{(g)}(S), \quad g \in \mathcal{G} \quad (9)$$

A positive value means the source helps the target under the declared game; a negative value is termed negative transfer. Classification requires the sign of the game-mean value and agreement of at least two of the three games. With 60 campaigns, the off-diagonal graph contains $60 \times 59 = 3,540$ directed pairs. Directionality is tested separately over the 1,770 unordered pairs by comparing the signs and ranks of $V_{\text{rem}}(s \rightarrow t)$ and $V_{\text{rem}}(t \rightarrow s)$; this distinguishes a symmetric compatibility interpretation from a target-dependent evidence-value interpretation. All three games use the same five recorded coordinates and therefore share the same blind spots: surface state, heater construction, detailed inlet development, metrology, and target-derived regime labels are unavailable.

2.9 Coalition value and confounding audit

Removal from the complete source pool measures a local marginal and can hide redundancy. Permutation Shapley values, therefore, average the utility increment over 64 random source orders for every target and transport game. Utility is a negative target MAPE relative to the positive Bowring anchor:

$$\phi_{s,t}^{(g)} = \mathbb{E}_{\pi}[U^{(g)}(S_{\pi,s} \cup \{s\}; t) - U^{(g)}(S_{\pi,s}; t)] \quad (10)$$

$$U^{(g)}(S; t) = -\text{MAPE}_t^{(g)}(S), \quad g \in \mathcal{G}$$

Monte Carlo standard errors and the game-specific Shapley efficiency residuals are archived. A ridge audit then regresses the game-mean removal value on standardized median distance, mean relative support, log source size, and log target size:

$$V_{\text{mean},s,t} \equiv \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} V_{\text{rem}}^{(g)}(s \rightarrow t) = \beta_0 + \boldsymbol{\beta}^T \mathbf{c}_{s,t} + \varepsilon_{s,t} \quad (11)$$

$$V_{\text{adj},s,t} = V_{\text{mean},s,t} - (\hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \mathbf{c}_{s,t})$$

A negative edge that becomes non-negative after adjustment is labeled nuisance-explained; a surviving negative residual is only a candidate for conflict. These categories identify replication priorities and are not evidence of causal facility bias. Shapley values are likewise conditional on the declared utility and learner. To quantify transport-rule sensitivity rather than average it away, the complete Shapley calculation is retained separately for each game, and source rankings are compared using Spearman's correlation. This tests whether precise coalition ordering is stable to continuous versus truncated notions of source support. A complete Shapley recomputation across unrelated learner families is not available and remains an explicit limitation.

2.9.1 Independent hierarchical response-field audit (TRACE)

To test whether the campaign result is robust to different estimators and target definitions, an independent hierarchical TRACE audit is conducted on the same set of 24,579 locked records and 60 RIDs. TRACE fits a multiscale radial-basis field to $\log[q\text{CHF}/(G \text{ hfg})]$ using only five independent pre-CHF coordinates, $\ln(P/P_{\text{crit}})$, $\ln(G/G_0)$, $\ln(D/D_0)$, $\ln(L/D)$, and Δhsub , $\ln/\text{hfg}$, with explicit campaign random intercepts and random slopes. The audit uses the same 12 whole-campaign outer folds and three independent seeds, but it is not used to select the primary champion, conformal width, transfer edges, or acquisition policy. Campaign identity is included only as a hierarchical random effect within TRACE; target-derived outlet quantities remain excluded from the design interface.

TRACE evaluates response topology for $K=1, \dots, 5$ using a predeclared joint rule: $\Delta\text{BIC} \geq 100$ relative to $K=1$, silhouette ≥ 0.20 , median campaign-bootstrap ARI ≥ 0.70 over 300 resamples, RBF-versus-local-estimator ARI ≥ 0.25 , minimum regime weight ≥ 0.05 , and support from at least five campaigns. Random-effect reproducibility is assessed using campaign-level Spearman correlations across seeds, after averaging fold-level estimates for each RID/seed. Intercept and slope magnitudes are summarized only after the same fold/seed aggregation. The labels R1, R2, ... are phenomenological response-field patterns, not DNB/dryout assignments. Because TRACE uses the same locked archive, agreement with the primary analysis demonstrates estimator independence rather than data independence.

2.10 Outcome-blind sparse assimilation and acquisition replay

Every outer-test campaign begins with its cross-fitted zero-shot log prediction. At $k \in \{1, 3, 5, 10\}$, TWIN-VALUE, maximin-coverage, uncertainty-driven, and repeated-random policies select candidate conditions using only information available before the selected CHF is revealed; random selection is repeated 12 times. The selected CHF is then revealed, the support row is removed from evaluation, and the remaining rows are re-predicted with the same regularized update. Fifty-eight campaigns contain enough query rows for the paired five-shot analysis. Two estimands are predeclared and kept separate at every budget: adaptation relative to zero-shot and acquisition-policy performance relative to repeated random selection.

The campaign correction uses an intercept-plus-standardized-feature basis $\mathbf{b}(\mathbf{x})$. For k revealed measurements, $\mathbf{B}_k$ stacks the k basis rows, and $\boldsymbol{\delta}_k$ contains the corresponding zero-shot log residuals. The feature-slope prior precision is 800, and the intercept prior precision is 80; the observation standard deviation is $\sigma\delta = 0.18$ in log space. Observed residuals are clipped to ± 0.25 around their median, and predicted corrections are limited to ± 0.80 :

$$\boldsymbol{\Sigma}_k = \left(\boldsymbol{\Lambda}_0 + \frac{\mathbf{B}_k^T \mathbf{B}_k}{\sigma_\delta^2} \right)^{-1}, \quad \boldsymbol{\mu}_k = \boldsymbol{\Sigma}_k \frac{\mathbf{B}_k^T \boldsymbol{\delta}_k}{\sigma_\delta^2} \quad (12)$$

$$\hat{q}_k(\mathbf{x}) = \hat{q}_0(\mathbf{x}) \exp\{\text{clip}[\mathbf{b}(\mathbf{x})^T \boldsymbol{\mu}_k, -0.80, 0.80]\}$$

2.10.1 Prior specification, shrinkage, and clipping rationale

The update is intentionally intercept-dominant. With only 1–10 measurements, a campaign-wide residual offset is identifiable more reliably than a high-dimensional rotation of the response surface. The tenfold larger precision on feature slopes therefore suppresses large condition-dependent corrections until the revealed data provide sufficient evidence. In comparison, the weaker intercept precision permits a facility-wide level shift. These precisions and $\sigma\delta$ are frozen regularization controls of the archived replay; they are not re-estimated from the held-out campaign and are not presented as universal Bayesian priors for CHF. Clipping provides a second fail-safe. The ± 0.25 residual window limits the leverage of an isolated revealed outlier on the fitted campaign state. At the same time, the ± 0.80 prediction correction confines the multiplicative adjustment to $\exp(-0.80)$ – $\exp(+0.80)$, approximately 0.45–2.23 times the zero-shot prediction. The main evidential test of these controls is the archived regularized-versus-unregularized comparison in the replay. A complete test-set-free sensitivity analysis of the prior precisions and clipping thresholds was not available in the archived analysis; this remaining hyperparameter dependence is stated in Section 5.

The TWIN-VALUE score is retained only as the name of the most elaborate outcome-blind acquisition policy. It combines geometric diversity, campaign distance, integrated posterior-variance reduction, value-weighted source disagreement, and ensemble uncertainty. All components are scaled by their 5th and 95th percentiles:

$$A_{TV}(\mathbf{x}) = 0.72D_{div}(\mathbf{x}) + 0.10D_{dom}(\mathbf{x}) + 0.08\Delta V_{post}(\mathbf{x}) + 0.05S_{src}(\mathbf{x}) + 0.05U_{ens}(\mathbf{x}) \quad (13)$$

The campaign-value prior for target t is learned only from transfer edges whose targets are not t . Consequently, no current target CHF can be included in the acquisition score. Adaptation relative to zero-shot and policy superiority relative to random are predeclared as distinct paired estimands. A successful Bayesian update does not validate the acquisition policy, and a failed acquisition-superiority gate does not invalidate the update itself.

2.11 Bounded thermal calculation and non-operational physics interface

For a uniformly heated round tube with constant mass flux and diameter, negligible axial wall conduction, no external heat loss, and pressure held fixed for property evaluation, the imposed heat flux adds enthalpy according to

$$h(z) - h_{f,sat} = -\Delta h_{sub,in} + \frac{4q''z}{GD}, \quad x_e(z) = \frac{h(z) - h_{f,sat}}{h_{fg}} = \frac{-\Delta h_{sub,in} + \frac{4q''z}{GD}}{h_{fg}} \quad (14)$$

The equation is evaluated at 51 axial nodes for 238 zero-shot and adapted states and at 0.8 and 1.0 times the predicted limiting heat flux. Numerical slopes are compared with the analytical value $4q''/(GD)$, and an independently archived relative conservation residual is required to be $\leq 10^{-10}$. This verifies only the numerical implementation of the stated steady energy balance. It does not validate pressure drop, wall temperature, void fraction, flow regime, interfacial transport, transient response, margin-to-CHF, or CHF location [66].

2.11.1 What the surrogate is not

The adaptive surrogate does not represent the evolving state of a reactor or test loop, synchronize with live sensors, solve momentum conservation, model two-phase pressure drop or void fraction, predict wall temperature, locate the CHF position, or forecast a transient approach to crisis. Selected historical CHF values update only a statistical residual state in log-prediction space. The energy calculation is therefore a conservation-based interface around a CHF surrogate, not a multiphase simulator and not an operational digital twin. The principal assumptions, numerical audits, and corresponding limits of inference of the present framework are summarized in Table 4.

Table 4 | Assumptions, numerical audits, and the resulting limits of inference.

Component	Assumption or condition	Audit or sensitivity	Permitted inference
Database	Published round-tube records and RID labels are correct	SHA-256 lock; completeness and positivity checks	Reproducible analysis of the archived evidence
Bowring anchor	Correlation evaluated in stated SI form	29 raw values at or below the 1 kW m ⁻² floor exported; exclusion sensitivity	Reference correlation only
Outer testing	RID approximates an experimental campaign	Disjoint role manifest; every RID is tested once	Transfer to unseen archived campaigns
Residual learning	Log discrepancy is learnable from pre-CHF inputs	Direct-model comparator and block-wise selection	Predictive correction, not mechanism discovery
Conformal intervals	Validation residuals are informative for outer campaigns	Three coverage estimands and interval width	Empirical marginal coverage only
Transfer value	Declared local experts and games define utility	Three games, sign audit, and Shapley comparison	Model-conditional source usefulness
Confounding adjustment	Four measured nuisances summarize observable support	Ridge residual and sign-change audit	Hypothesis ranking, not causal diagnosis
Bayesian update	An intercept dominates the sparse campaign residual	Strong slope shrinkage; unregularized comparator	Stable retrospective sparse residual correction; not state synchronization

Component	Assumption or condition	Audit or sensitivity	Permitted inference
Acquisition	Candidate coordinates and diagnostics are available before CHF	No hidden-label input; support manifest	Outcome-blind policy comparison; no optimality claim
Energy interface	Steady 1-D balance; fixed G, D and pressure; uniform q''	Analytical closure and residual $\leq 10^{-10}$	Energy conservation only; not CFD
Prospective matrix	Candidates remain within the observed input envelope	Blind selection flag and frozen ranking	Frozen proposed falsification experiment; not prospective validation

2.12 Predeclared claim gates, software, and reproducibility

Before the full analysis, 17 numerical decision gates were fixed to prevent the narrative from being optimized after inspecting test outcomes (Table 5). Eleven gates are critical, and all eleven pass. Two non-critical claims fail and remain part of the evidence: G07C reaches a 24.4% five-query gain rather than its 25% stretch target, and G08B does not establish TWIN-VALUE superiority over repeated random selection. The full primary run uses seed 20260814, 12 outer blocks, 320 trees, 64 Shapley permutations, 4,000 campaign-bootstrap replicates, 12 random-policy repeats, and support sizes 0, 1, 3, 5, and 10. The archived environment records Python 3.13.5, NumPy 2.5.2, pandas 2.3.3, SciPy 1.18.0, and scikit-learn 1.9.0. The independent hierarchical response-topology audit uses the same 12 campaign-held-out folds, three seeds, and a separately archived software environment.

Table 5 | Predeclared claim ledger. Critical, important, and exploratory decisions are stated with numerical operators and thresholds; failed gates remain visible and constrain the claims that follow.

Gate	Severity	Statement	Observed	Threshold	Status
G01	Critical	All 60 campaigns receive exactly one outer test assignment	60	= 60	Pass
G02	Critical	Champion record MAPE $\leq 5.5\%$	4.517%	$\leq 5.5\%$	Pass
G03	Critical	Champion improves record MAPE by at least 25% versus Bowring	53.3%	$\geq 25\%$	Pass
G04	Critical	Champion improves campaign-macro MAPE by at least 25% versus Bowring	53.5%	$\geq 25\%$	Pass
G05	Important	Champion campaign p_{90} MAPE $\leq 20\%$	16.31%	$\leq 20\%$	Pass
G06	Critical	The acquisition routine contains no hidden-label input	True	= True	Pass
G06B	Critical	Cross-target campaign-value prior excludes current target labels	True	= True	Pass
G07A	Critical	Five selected measurements reduce adaptive-surrogate macro MAPE by $\geq 5\%$	24.4%	$\geq 5\%$	Pass
G07B	Critical	Paired campaign-bootstrap 95% lower bound for $k=5$ gain is positive	+0.515	> 0	Pass
G07C	Important	Stretch gate: five selected measurements reduce adaptive-surrogate macro MAPE by $\geq 25\%$	24.4%	$\geq 25\%$	Fail

Gate	Severity	Statement	Observed	Threshold	Status
G08A	Important	TWIN-VALUE acquisition at $k=5$ is within 0.25 MAPE points of repeated random selection	0.051	≤ 0.25	Pass
G08B	Exploratory	Paired bootstrap lower bound shows TWIN-VALUE acquisition beats repeated random selection	-0.305	> 0	Fail
G09	Critical	Maximum Shapley efficiency residual $\leq 10^{-8}$ MAPE points	3.38×10^{-14}	$\leq 10^{-8}$	Pass
G10	Critical	Maximum relative energy conservation residual $\leq 10^{-10}$	3.07×10^{-12}	$\leq 10^{-10}$	Pass
G11	Important	At least 70% of directed edges have a unanimous/majority sign across games	94.3%	$\geq 70\%$	Pass
G12	Critical	Campaign-robust 95% interval achieves $\geq 90\%$ campaign-macro coverage	92.17%	$\geq 90\%$	Pass
G13	Important	Campaign-robust 95% median upper/lower width ratio ≤ 4	3.629	≤ 4	Pass

2.13 Integrated adaptive-surrogate workflow

Figure 1 summarizes the separation between cross-campaign learning, sparse residual updating, and thermal verification. Reference campaigns create the frozen zero-shot model and target-excluding campaign-value prior; measurements from a held-out campaign are then revealed sequentially and used only for the regularized Bayesian residual update and the declared acquisition comparison. The energy balance verifies conservation separately. The TWIN-VALUE name appearing in the archived workflow figure denotes this analysis workflow/acquisition score and should not be read as an operational digital-twin classification. Fig. 1 presents the integrated campaign-aware workflow for prediction, transfer, adaptation, and verification.

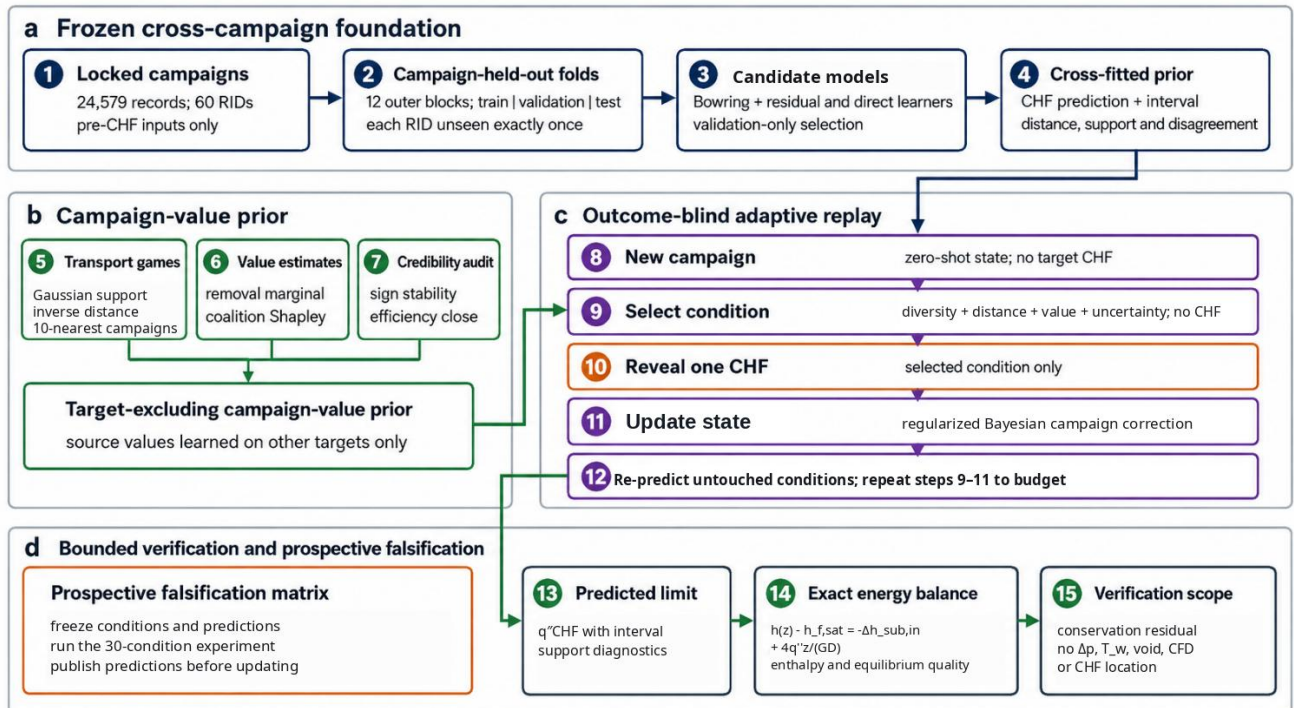


Fig. 1 | Campaign-aware framework for CHF prediction, transfer, and adaptation. (a) Frozen campaign-held-out foundation; (b) target-excluding campaign-value prior; (c) outcome-blind measurement selection followed by sparse residual updating; and (d) bounded energy verification and prospective test design. Target campaign CHF enters only after a query condition is selected and revealed.

2.14 Prospective commitment and falsification boundary

The 30-condition matrix is frozen only for the retrospective analysis: its candidate coordinates, zero-shot predictions, intervals, and priority ranking must remain unchanged before any future measurements are inspected. A valid prospective test would require an independent experimental campaign, complete surface and metrology documentation, a precommitted comparison protocol, and publication of null or adverse outcomes without refitting the frozen predictions. The current manuscript does not claim that such an external experiment has been executed.

No external preregistration was created for the prospective matrix. Figure 16 is therefore treated as a versioned future test plan rather than as prospective evidence, and it does not strengthen the current retrospective claims of accuracy, coverage, transfer, or adaptation.

3. Results

The results are organized around a single evidentiary question: what survives when the experimental campaign, not the database row, is treated as the unit of generalization? Prediction accuracy is the first instrument, not the endpoint. We first separate historical record concentration from unequal campaign difficulty, then test physics-anchored prediction on wholly unseen experiments. We next ask whether the same structure appears in calibration and applicability; whether one experiment helps another symmetrically or directionally; whether coalition rankings survive changes in the transport game; whether an estimator-independent hierarchical analysis detects a reproducible campaign state; and whether a few revealed measurements can recover that state safely. The thermal calculation and prospective matrix only define the boundary and falsifiability of the offline surrogate.

3.1 Experimental provenance changes the CHF generalization problem

The campaign-held-out protocol makes the deployment question explicit. Across 12 outer blocks, each of the 60 campaigns appears exactly once as unseen test evidence, while preprocessing, model fitting, candidate selection, ensemble weighting, interval calibration, and acquisition priors remain confined to disjoint training and validation campaigns (Fig. 2a,b).

All 11 predeclared critical claim gates pass (Fig. 2c; Table 5), while two non-critical gates constrain interpretation. G07C misses the 25% five-measurement stretch target by 0.6 percentage points, and G08B fails because the 95% paired interval for TWIN-VALUE versus repeated random acquisition crosses zero. The results, therefore, support sparse updating but not the superiority of the acquisition policy.

The information firewall, therefore, fixes the evidentiary boundary for all subsequent comparisons: predictive accuracy, uncertainty, source value, and adaptation are evaluated only after the target campaign has been excluded from the decisions that define them. Fig. 2 shows the campaign-resolved validation framework and outcomes of the predeclared claim tests.

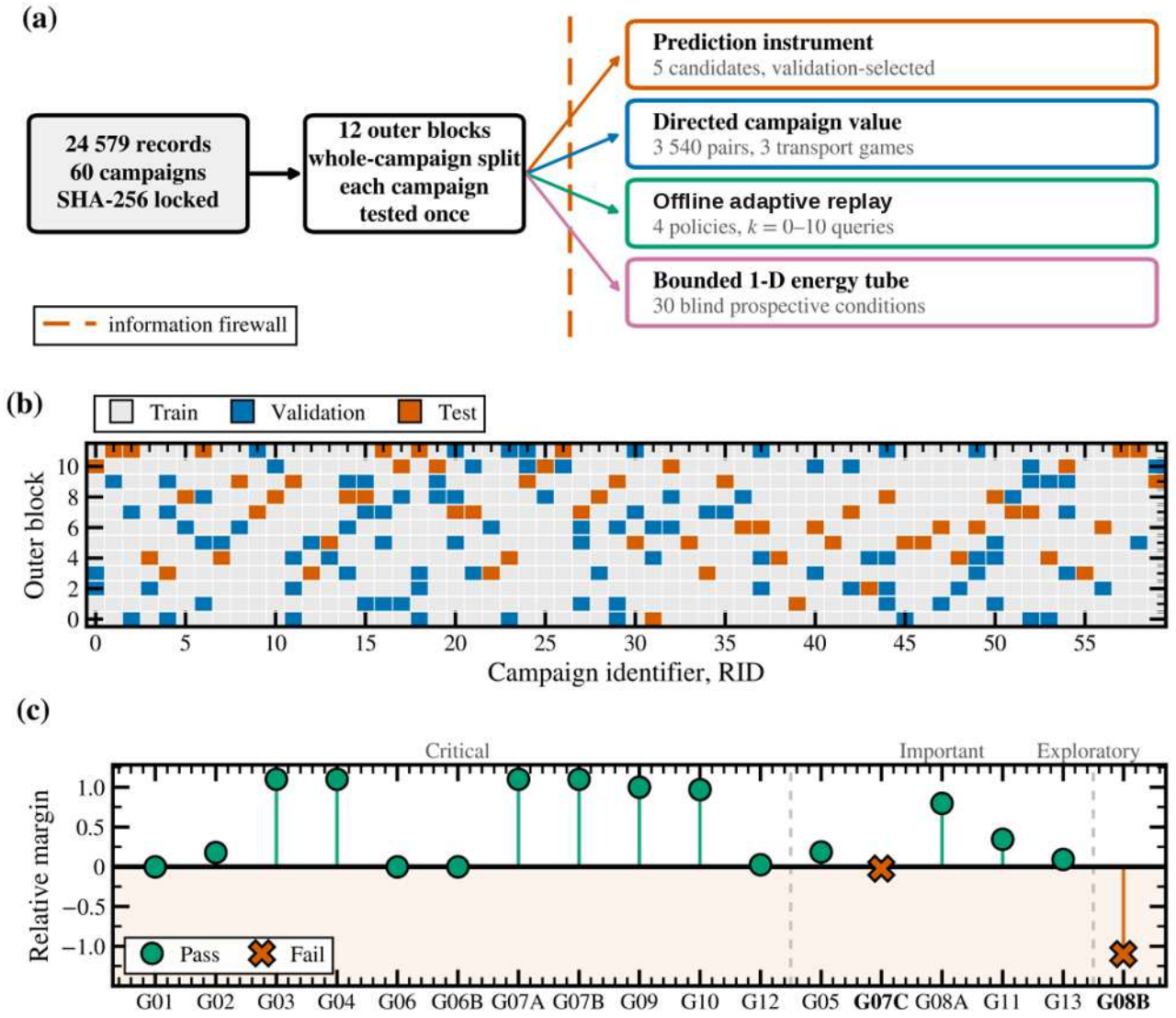


Fig. 2 | Campaign-resolved validation and claim testing. (a) Data flow through the 12 whole-campaign outer blocks; (b) train, validation, and test roles assigned to the 60 campaigns; and (c) outcomes of the predeclared numerical gates. Each campaign serves as a completely unseen test of evidence.

The need for campaign-level separation becomes apparent from the structure of the evidence itself. The 24,579 records span a broad pressure–mass-flux envelope, but individual campaigns occupy much narrower trajectories through that space (Fig. 3a,b). Diameter occurs in discrete manufactured bands, particularly around 5, 8, and 10 mm, and the maximum L/D approaches 2,500.

The distribution of records among experiments is strongly concentrated: the five largest campaigns contain 54.0% of all measurements, whereas the thirty smallest contain only 5.9%, giving a record-concentration Gini coefficient of 0.719 (Fig. 3c). The median campaign spans only 18% of the global logarithmic pressure range (Fig. 3d). This concentration is a property of how the historical archive was assembled, not a physics result; its role is to explain why equal-record and equal-campaign estimands answer different questions.

These observations establish the statistical premise for the remainder of the study. A campaign is an experimental trajectory with shared geometry, instrumentation, surface history, and operating practice, while the archive gives some trajectories thousands of times more weight than others. Record-weighted evaluation, therefore, emphasizes experiments with dense sampling. Campaign-macro evaluation deliberately removes that historical sampling advantage and asks about transfer to a typical physically separate experiment. The two estimands are complementary rather than

competing measures. The thermohydraulic coverage and sampling imbalance of the CHF archive are shown in Fig. 3.

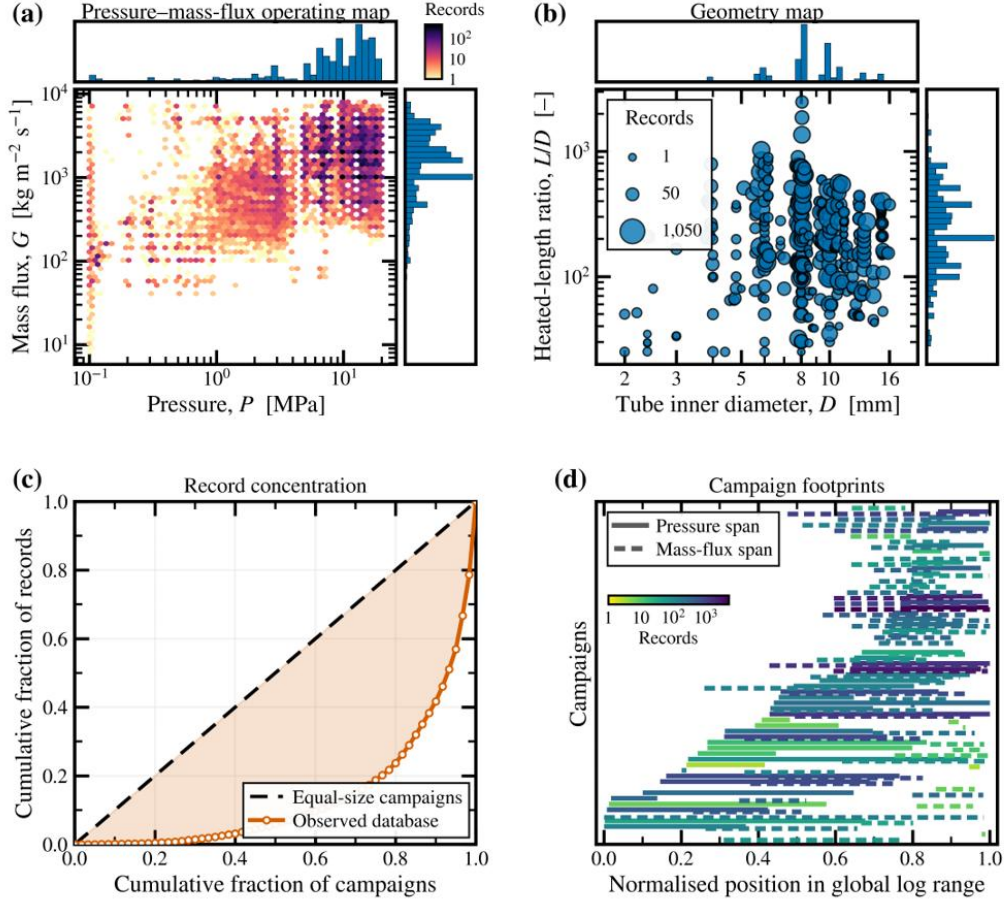


Fig. 3 | Thermohydraulic coverage and sampling imbalance of the CHF archive. (a) Pressure–mass-flux operating space; (b) diameter and heated-length-ratio coverage; (c) concentration of records among campaigns; and (d) campaign-wise operating spans relative to the full database. Global coverage is broad, but individual campaigns are much narrower and unevenly sampled.

3.2 Physics anchoring improves prediction, but does not remove campaign heterogeneity

Having established the campaign as the relevant transport unit, we next asked whether the physics-anchored learner serves as a useful predictive instrument under that stricter test. Against the fixed Bowring correlation, the validation-selected champion reduces record-weighted MAPE from 9.67% to 4.52%, removes the mean signed bias to approximately zero, and increases the fraction of records predicted within $\pm 20\%$ from 87.3% to 97.4% (Fig. 4a,b; Table 6).

The improvement is not uniform across the CHF range. Across fourteen CHF bins, the Bowring median signed error exceeds $\pm 5\%$ in seven bins, compared with three for the champion. In contrast, the largest median deviation decreases from -36.8% to -11.7% (Fig. 4c). The strongest anchor deficiencies occur below approximately 0.3 MW m^{-2} and above approximately 5 MW m^{-2} . These regions identify limitations of the fixed closure within the available evidence; they cannot be interpreted as newly detected DNB or dryout mechanisms because the archive contains no corresponding wall-temperature histories, flow visualization, or film-thickness measurements.

The improvement is also distributional rather than being produced by a small number of large corrections: 17,833 of 24,579 records (72.6%) are closer to the measurement under the champion, and its error distribution remains below that of Bowring across the observed quantiles (Fig. 4d). Thus, the learned discrepancy correction is a substantially more accurate predictive instrument, while the

residual structure remains a diagnostic of where the physical anchor transfers least reliably. Cross-fitted CHF prediction relative to the Bowring baseline is presented in Fig. 4.

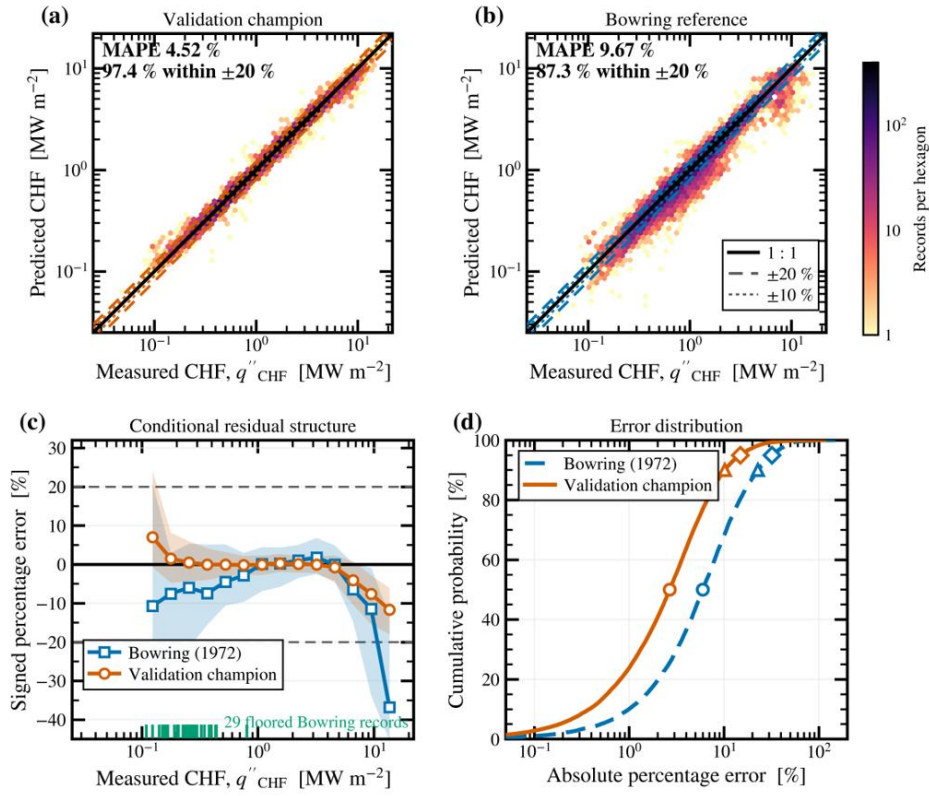


Fig. 4 | Cross-fitted CHF prediction relative to Bowring. (a) Measured versus surrogate-predicted CHF; (b) measured versus Bowring-predicted CHF; (c) median signed error across the CHF range; and (d) absolute-error distributions. The physics-anchored surrogate reduces systematic and distributional error across held-out campaigns.

More consequentially, the apparent magnitude of prediction error varies across reporting units. For the champion, MAPE rises from 4.517% under equal record weighting to 7.474% when every campaign receives one vote and reaches 32.929% for the most difficult campaign (Fig. 5a; Table 6). The campaign-macro/record-weighted ratio is 1.65 for the champion and 1.66 for Bowring; across all seven predictors, it remains between $\times 1.65$ and $\times 1.69$ (Fig. 5b). This near-invariance makes the reporting-unit effect an attribute of the evidence structure rather than of one architecture. Still, it does not by itself tell whether the effect comes from record concentration or unequal campaign difficulty.

The weight-only decomposition answers that narrower question without requiring counterfactual retraining. Within campaign-size terciles, the macro/record ratios are only $\times 1.05$, $\times 1.11$, and $\times 1.31$, whereas the full archive ratio is $\times 1.65$. Capping each campaign at 500 records while leaving each prediction unchanged increases the weighted MAPE from 4.517% to 5.676%. Thus, 39.2% of the 4.517→7.474 MAPE gap is recovered by reweighting alone, and 60.8% remains as unequal campaign-difficulty heterogeneity (Section 3.2.1; Table 7). The latter is deliberately not labeled a physics fraction because facility state, support, metrology, and omitted variables can all contribute.

The paired campaign analysis reaches the same conclusion without relying on an average. The champion improves on Bowring in 52 of 60 unseen campaigns, and campaign p90 MAPE decreases from 37.456% to 16.312% (Fig. 5c,d; Table 6). The improvement is broad, but the eight adverse campaigns remain visible and become the critical cases for asking what accuracy alone cannot explain. The contrast between record-level and campaign-level model performance is shown in Fig. 5.

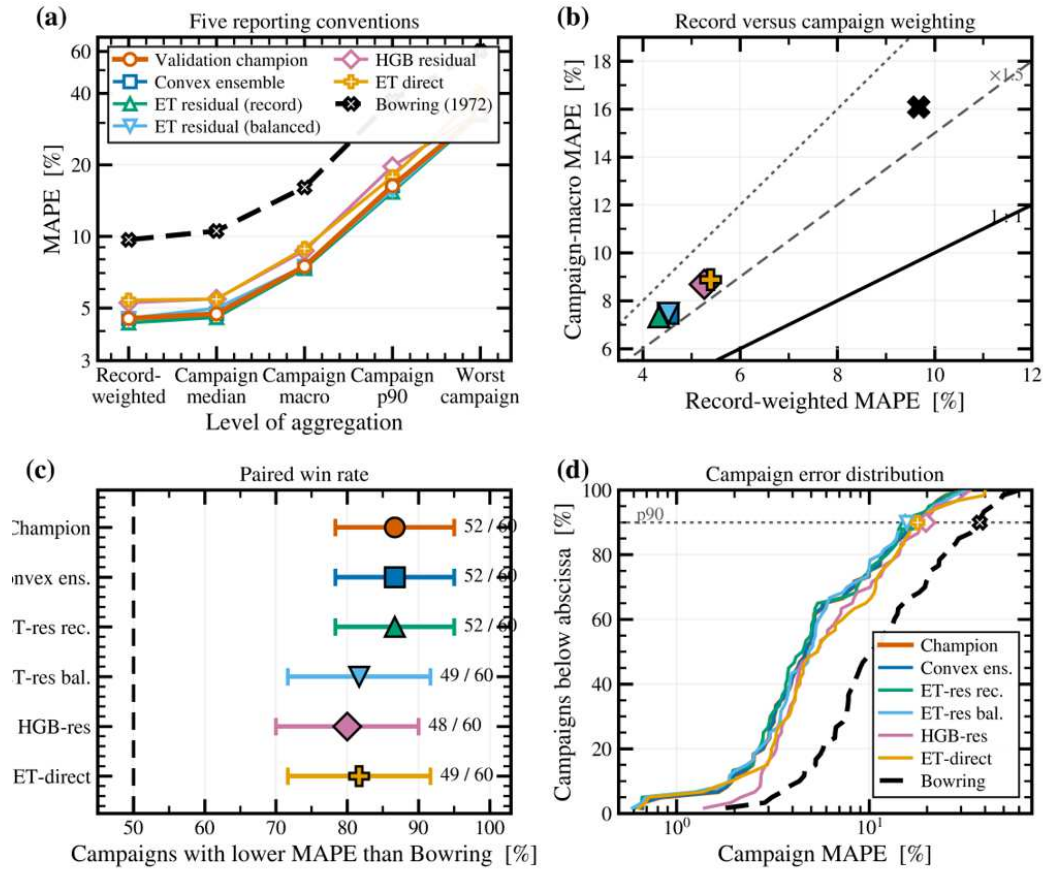


Fig. 5 | Record-level and campaign-level model performance. (a) Candidate-model performance under different aggregation rules; (b) record-weighted versus campaign-macro MAPE; (c) fraction of campaigns improved relative to Bowring; and (d) campaign-error distributions. Campaign-level evaluation consistently reveals greater difficulty with generalization than record-weighted reporting.

Table 6 | Cross-fitted predictive performance under record- and campaign-level reporting conventions.

Estimand	Evaluation unit	Bowring	Validation champion	Change
MAPE	All records, equal record weight	9.670%	4.517%	−53.3%
Median campaign MAPE	Median of 60 within-campaign MAPEs	10.530%	4.724%	−55.1%
Campaign-macro MAPE	Mean of 60 within-campaign MAPEs	16.084%	7.474%	−53.5%
Campaign p ₉₀ MAPE	90th percentile across campaigns	37.456%	16.312%	−56.5%
Worst-campaign MAPE	Maximum across campaigns	60.328%	32.929%	−45.4%
Signed bias	Mean signed percentage error over records	−2.265%	+0.036%	Near-zero bias
Within ±10%	Fraction of records	68.0%	89.6%	+21.6 points
Within ±20%	Fraction of records	87.3%	97.4%	+10.1 points
Campaigns improved	Paired campaign comparison	Reference	52 of 60	8 adverse campaigns
Records closer	Paired record comparison	Reference	17,833 (72.6%)	Champion closer

3.2.1 Historical sampling concentration explains only part of the campaign penalty

The size-tercile audit separates two effects that are otherwise confounded in the headline $\times 1.65$ ratio. The 20 largest campaigns account for 21,381 of 24,579 records and have an equal-campaign MAPE of 5.243%, whereas the middle and smallest terciles have equal-campaign MAPEs of 8.822% and 8.358%, respectively. Dense historical campaigns are therefore both heavily weighted and, on average, easier for the frozen champion. Within each size tercile, however, record and campaign weighting are much closer than at the archive level (Table 7).

The 500-record cap provides a direct weight-only perturbation. Without retraining or changing any prediction, weighted MAPE moves from 4.517% to 5.676%. Relative to the 7.474% campaign-macro endpoint, this reweighting closes 39.2% of the gap, leaving 60.8% unresolved. The defensible conclusion is therefore not that 60.8% is 'physics', but that historical row counts alone cannot explain the campaign penalty. Most of the remaining difference is attributable to which experiments are difficult, and to whatever combination of support, apparatus, metrology, surface state, and unrecorded physics produces that difficulty.

Table 7 | Weight-only decomposition of the champion record-to-campaign MAPE gap into historical sampling concentration and residual campaign-difficulty heterogeneity. Predictions are frozen; no model is retrained.

Stratum	Campaigns	Records	Record-weighted MAPE	Equal-campaign MAPE	Ratio
Smallest size tercile	20	475	7.945%	8.358%	$\times 1.05$
Middle-sized tercile	20	2,723	7.947%	8.822%	$\times 1.11$
Largest size tercile	20	21,381	4.004%	5.243%	$\times 1.31$
All campaigns	60	24,579	4.517%	7.474%	$\times 1.65$
All campaigns; cap at 500 records	60	24,579	5.676%	7.474%	$\times 1.32$

Resolving all 60 held-out experiments individually shows that the adverse cases are not confined to one learner family (Fig. 6a,b). Campaigns that are difficult for the champion also tend to be difficult for other candidates, resulting in a predominantly row-wise, rather than model-specific, structure in the campaign-by-model error map (Fig. 6c).

This shared difficulty supports a campaign-level missing-state interpretation but does not identify its cause. The available metadata cannot separate limited thermohydraulic support, surface or heater condition, facility metrology, inlet development, response-pattern composition, or interactions among these factors. Campaign provenance is therefore an evidentiary state variable, not a thermophysical input variable.

Campaign size alone does not explain the paired gains. Improvement over Bowring decreases with campaign size (Spearman $\rho = -0.381$, $p = 0.003$). Yet, the 12 campaigns containing at least 500 records retain a median improvement of +5.44 MAPE points (Fig. 6d). Together with Table 7, this shows that historical concentration matters but cannot fully account for the difficulty pattern in unseen campaigns. Campaign-wise generalization across the unseen experiments is detailed in Fig. 6.

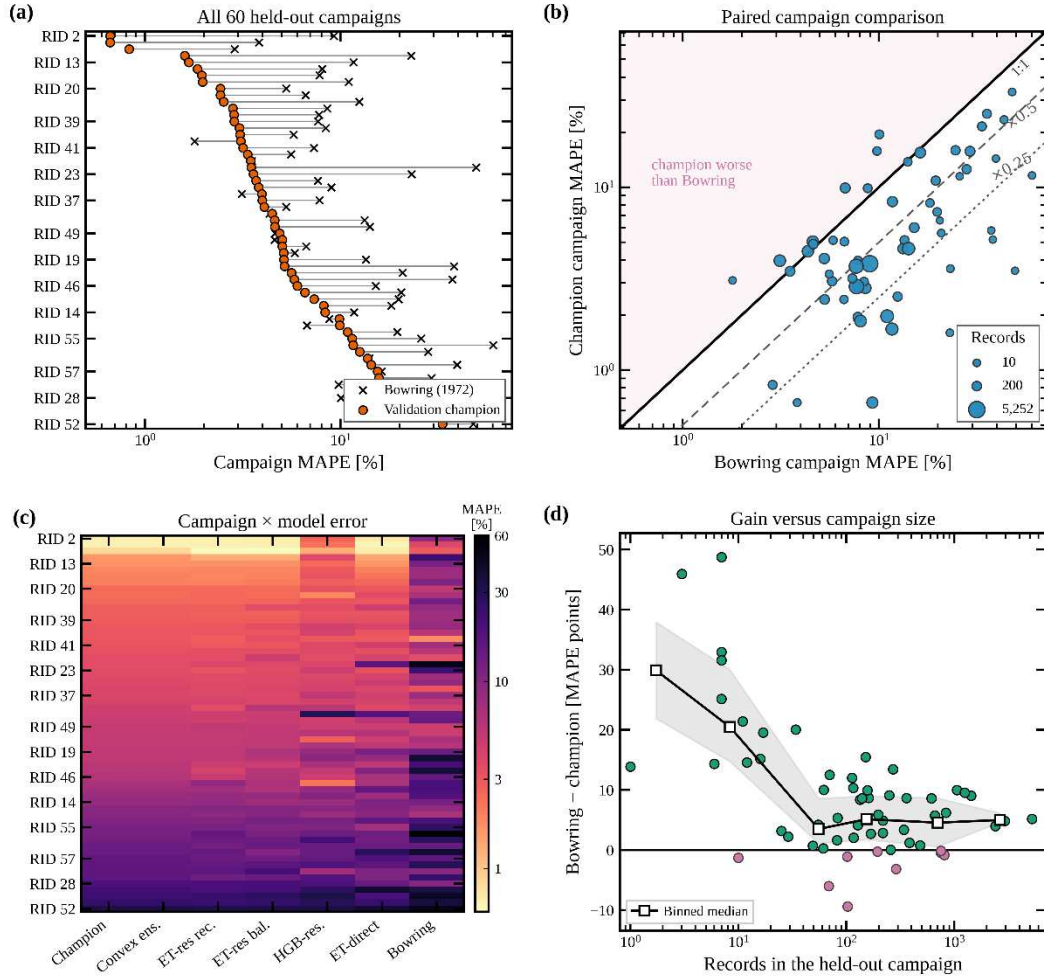


Fig. 6 | Generalization across individual unseen campaigns. (a) Campaign-wise Bowring and surrogate errors; (b) paired campaign comparison; (c) campaign difficulty across model families; and (d) MAPE improvement versus campaign size. Most campaigns improve, but several remain systematically difficult across different learners.

The information firewall also imposes a measurable model-selection cost. Candidate performance varies across the 12 outer blocks, and the validation-fitted convex weights vary with it (Fig. 7a,b); nevertheless, the selector chooses the convex ensemble in 11 of 12 blocks, with no single learner assumed to dominate every target campaign.

Relative to the retrospectively identified test-best candidate, the validation-selected model incurs a mean regret of 0.31 MAPE points and a maximum regret of 1.20 points, against a within-block candidate spread of up to 5.33 points (Fig. 7c; Table 8). The selector never chooses the test-worst candidate. Figure 7d is a separate calibration-cost audit: campaign-robust conformal half-widths are $\times 1.7$ – $\times 6.9$ wider than pooled half-widths across blocks, with a median factor of $\times 2.5$.

Only 5.8% of the largest exploitable within-block candidate spread is lost on average by enforcing validation-only selection. The residual campaign heterogeneity, therefore, cannot reasonably be dismissed as a test-set selection artifact. Fig. 4, Fig. 5, Fig. 6 and Fig. 7 establish the first central result: physics-anchored discrepancy learning markedly improves CHF prediction across unseen experiments, but honest selection and model-family choice do not remove the experiment-level component of difficulty. Validation-only model selection and its calibration cost are reported in Fig. 7.

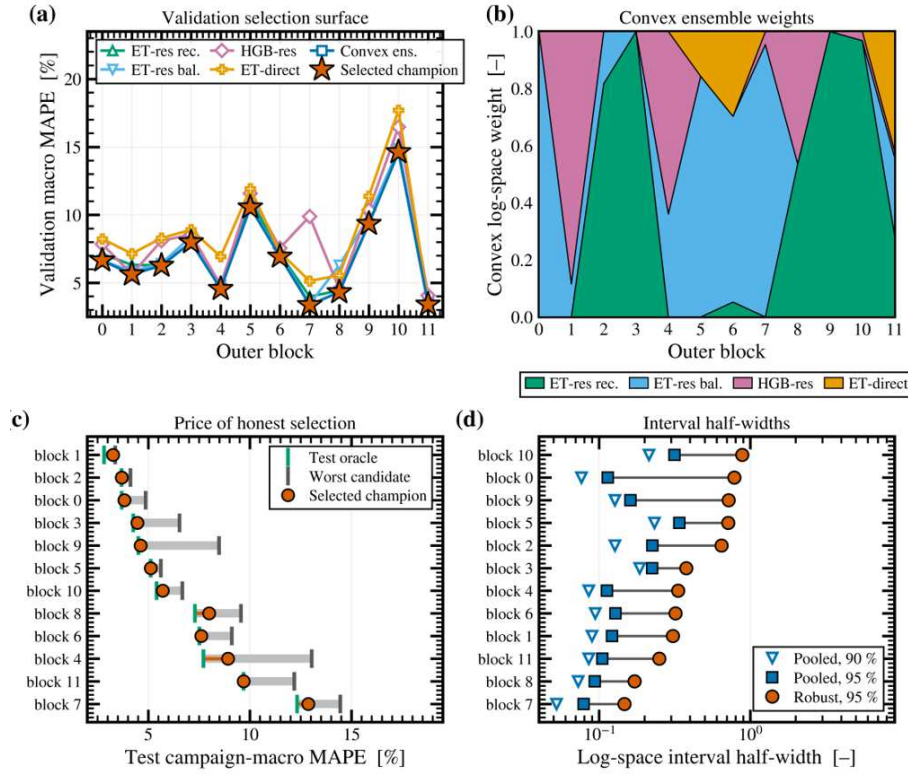


Fig. 7 | Validation-only model selection and calibration cost. (a) Validation performance of the candidate models; (b) block-wise convex-ensemble weights; (c) selected-model regret relative to the retrospectively identified test-best candidate; and (d) pooled versus campaign-robust conformal half-widths. Honest selection adds little regret, whereas stronger campaign protection requires wider intervals.

Table 8 | Validation-only selection audit and retrospective calibration-width cost relative to unavailable test information.

Audit quantity	Definition	Observed value	Interpretation
Mean selection regret	Selected test MAPE – test-best candidate MAPE	0.31 MAPE points	Average cost of preserving the firewall
Maximum selection regret	Largest block-wise regret	1.20 points	Upper observed selection cost
Candidate spread	Largest within-block test-MAPE range	5.33 points	Model choice matters
Worst candidate selected	Blocks selecting the test-worst candidate	0 of 12	No catastrophic selection
Convex ensemble selected	Blocks minimizing validation macro MAPE	11 of 12	Ensemble usually preferred
Mean convex weights	ET residual A / balanced / HGB residual/direct	0.39 / 0.37 / 0.17 / 0.07	Selection remains distributed across model families
Relative regret	Mean regret / largest exploitable spread	5.8%	Most exploitable variation retained
Robust/pool half-width ratio	Across 12 outer blocks	$\times 1.7$ – $\times 6.9$; median $\times 2.5$	Retrospective cost of campaign-robust calibration

3.3 Campaign shift also governs uncertainty and pre-measurement applicability

If campaign provenance affects prediction error, a deployment-relevant uncertainty estimate should expose the same structure. At nominal 95% coverage, pooled calibration reaches 91.85% when records are weighted equally, but only 82.49% when campaigns are weighted equally and 34.92% in the lower-decile campaign tail

(Fig. 8a; Table 9). The campaign-robust rule raises these quantities to 98.15%, 92.17%, and 76.99%, respectively.

The difference is visible at the individual-experiment level. Twenty of 60 campaigns fall below 90% coverage under pooled calibration versus 12 under the campaign-robust rule, and 10 campaigns are repaired to at least nominal coverage (Fig. 8b,c). Protection is not free: median within-campaign width inflates by $\times 1.47$, while the across-block half-width penalty varies much more strongly ($\times 1.7$ – $\times 6.9$; median $\times 2.5$; Fig. 7d). The median robust upper/lower multiplicative ratio is 3.629.

Coverage decreases as campaign prediction difficulty increases (Spearman $\rho = -0.642$; Fig. 8d). Satisfactory pooled-record coverage therefore does not imply comparable protection for a difficult unseen facility. The 92.17% robust macro coverage is a retrospective empirical property of the archived held-out experiments, not a theorem guaranteeing that rate for an arbitrary future facility with unrepresented surface, instrumentation, or response characteristics. Campaign-dependent uncertainty under pooled and campaign-robust calibration is compared in Fig. 8.

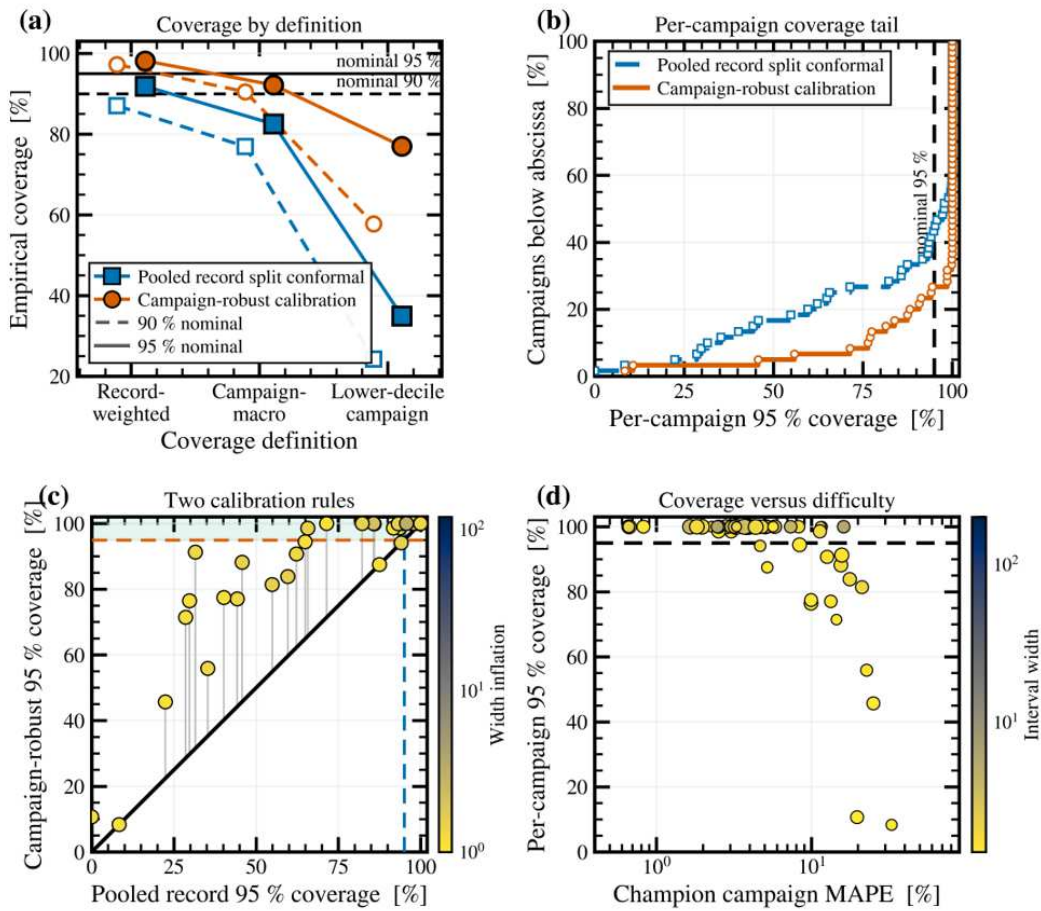


Fig. 8 | Campaign-dependent uncertainty under pooled and robust calibration. (a) Empirical coverage at record, campaign, and lower-tail levels; (b) distribution of 95% coverage across campaigns; (c) paired pooled and campaign-robust coverage; and (d) coverage versus campaign prediction error. Pooled calibration can look adequate at the record level while leaving difficult campaigns poorly protected.

3.3.1 Coverage is facility-specific, not universal

This distinction changes how the intervals should be used. A new campaign should first be checked against the archived support diagnostics. Its earliest revealed measurements should be compared with the frozen interval predictions before those intervals are treated as operationally informative. Persistent out-of-interval residuals, unusually large evidence distance, or model disagreement are reasons to widen, recalibrate, or withhold the surrogate rather than to invoke the retrospective 92.17% figure as a guarantee. No fixed safety factor or licensing threshold is inferred from the present database.

Table 9 | Empirical nominal-95% interval calibration under pooled-record and campaign-robust rules.

Coverage estimand	Evaluation unit	Pooled calibration	Campaign-robust calibration	Implication
Record-weighted coverage	Individual records	91.85%	98.15%	Pooled rule misses the nominal 95%
Campaign-macro coverage	Equal weight per campaign	82.49%	92.17%	Robust rule passes G12.
Lower-decile coverage	10th percentile of campaign coverage	34.92%	76.99%	Tail protection remains incomplete.
Campaigns below 90%	Count among 60	20	12	Eight fewer poorly covered campaigns
Median interval ratio	Upper/lower multiplicative width	1.276	3.629	Robustness costs width
Paired width inflation	Median within-campaign ratio	Reference	$\times 1.47$	Empirical calibration cost
Campaigns repaired	Moved to $\geq 95\%$ coverage	Reference	10 of 60	Partial facility-level repair

The same difficulty can partly be detected before CHF is measured. Absolute prediction error increases with standardized distance from the nearest training campaign ($\rho=+0.165$) and more strongly with candidate-ensemble disagreement ($\rho=+0.350$) (Fig. 9a,b). Both diagnostics are available from the query condition and frozen models alone.

When predictions are ranked by evidence distance, retaining the closest 50% reduces record-weighted MAPE from 4.52% to 3.44% without consulting the measured target CHF (Fig. 9c). The distance-decile audit in Fig. 9d shows the same qualitative degradation in accuracy and pooled coverage toward weakly supported regions. At the same time, the campaign-robust rule is more stable. Distance is therefore a warning variable rather than a certified applicability threshold.

Prediction error, conformal coverage, and outcome-blind support diagnostics thus converge on one practical conclusion: proximity to represented evidence matters, but geometric distance alone is insufficient. This motivates a directed transfer analysis that asks which historical experiments improve a target and, crucially, whether the reciprocal transfer has the same sign. The relationship between evidence support and prediction reliability is examined in Fig. 9.

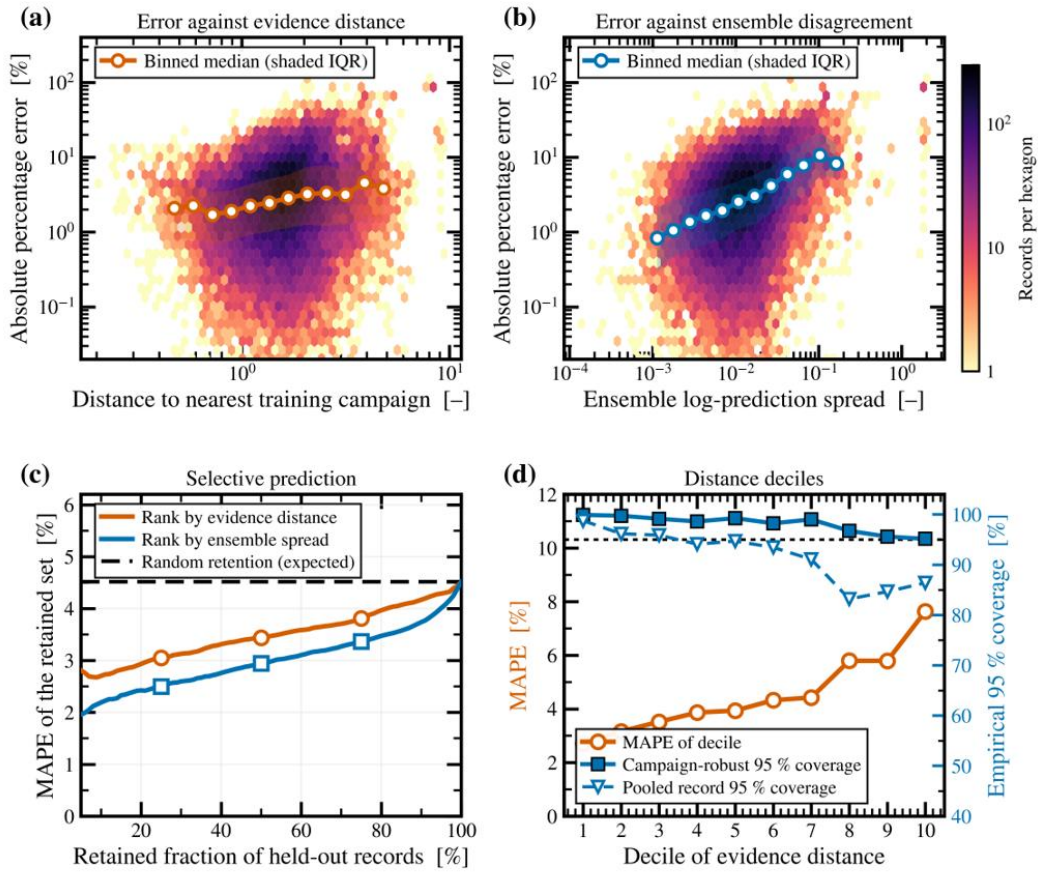


Fig. 9 | Evidence support and prediction reliability. (a) Absolute error versus distance from the training evidence; (b) error versus disagreement among candidate models; (c) risk–coverage behavior as less-supported predictions are retained; and (d) error and coverage across evidence-distance deciles. Reliability deteriorates as conditions move away from well-supported historical evidence.

3.4 Campaign transfer is directional and frequently harmful; measured support explains only part of the structure

Campaign information was next treated as a transferable scientific resource rather than an undifferentiated training pool. Across all 3,540 ordered source–target pairs, 1,537 transfers are helpful (43.4%), 1,802 harmful (50.9%), and 201 game-dependent (5.7%) under the declared transport games (Fig. 10a; Table 10). Most individual effects are small, the median absolute removal value is 0.042 MAPE points, but their sign is reproducible: 806 edges are unanimous, and at least two of three games agree for 3,339 edges (94.3%; Fig. 10b).

The negative-transfer structure follows support geometry. Harmful share rises from 31% in the closest distance stratum to 64% in the most distant and falls from 68% to 27% as relative source support increases (Fig. 10c,d). Thus, simply adding historical records is not guaranteed to improve a new experiment; value depends on whether the source actually supports the target operating region.

Distance and support do not determine the cause of a harmful edge. The archive cannot separate missing operating variables, facility/metrology effects, heater or surface history, residual bias in the physical anchor, and genuine changes in the local CHF response. The transfer map is therefore interpreted as predictive compatibility under a declared game, never as a mechanism map. Directed campaign-to-campaign transfer of predictive information is mapped in Fig. 10.

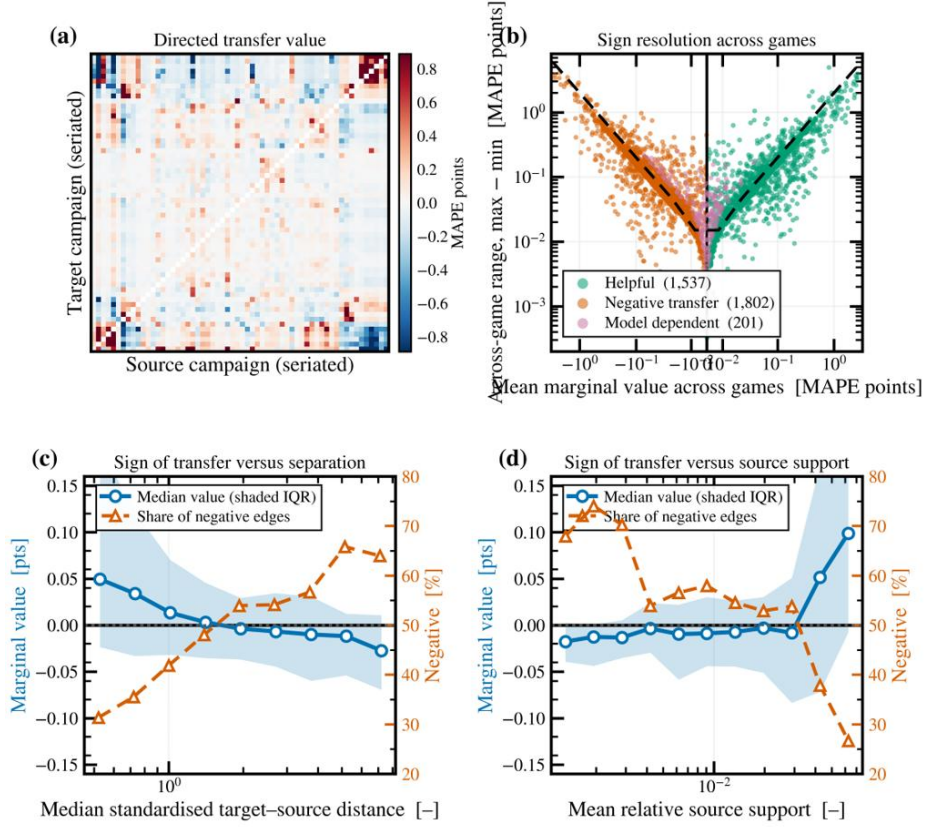


Fig. 10 | Directed transfer of predictive information between campaigns. (a) Source-to-target removal-value map; (b) transfer magnitude and sign consistency across the three transport games; (c) harmful-transfer fraction with increasing thermohydraulic distance; and (d) harmful-transfer fraction with relative source support. Historical campaigns are not interchangeable; their usefulness depends on the target direction and the support they receive.

3.4.1 Experimental value is directional rather than pairwise symmetric

The ordered graph supports a test that record-level analysis cannot perform. Across the 1,770 unordered campaign pairs, the two directions agree in sign for 70.0% and disagree for 30.0%; the rank association between $V_{rem}(s \rightarrow t)$ and $V_{rem}(t \rightarrow s)$ is only moderate (Spearman $\rho = +0.565$). In nearly one-third of campaign pairs, one experiment improves prediction for the other while the reciprocal transfer is harmful. Experimental value is therefore a property of the ordered source $\rightarrow$ target relation, not an intrinsic compatibility label attached to the pair.

This asymmetry is consistent with support geometry without proving a physical cause. A broad, densely sampled source can support a narrower target embedded within its operating region, whereas the narrow campaign may contribute little or detrimental information in the reverse direction. Any downstream use of transfer edges, to prioritize replication, down-weight a source, or flag a target, must therefore preserve direction and should never translate a one-way negative edge into a statement that two facilities are mutually inconsistent.

3.4.2 Measured support explains part, but not all, of the harmful transfer

The nuisance audit narrows the unresolved part of the graph. Adjustment for campaign size, thermohydraulic distance, and relative support renders 469 of the 1,802 majority-supported harmful edges non-negative (26%) and changes the sign of 550 edges overall (Fig. 11a,b; Table 10). The adjustment acts mainly near zero; the strongly negative tail persists. Candidate covariate-shift edges carry the largest median adjusted severity, 0.077 MAPE points, while distance retains a negative association with value ($\rho = -0.194$) and relative support a positive one ($\rho = +0.161$) (Fig. 11c,d).

The surviving negative tail is consequently a replication target rather than a causal diagnosis. Incomplete measured support, unrecorded apparatus/metrology effects, and genuine local response differences remain

simultaneously plausible. The independent TRACE analysis in Section 3.7 tests whether a reproducible response-field structure exists, but even that second estimator cannot uniquely assign the harmful edges to a single physical mechanism. The influence of measured factors and nuisance adjustment on harmful transfer is assessed in Fig. 11.

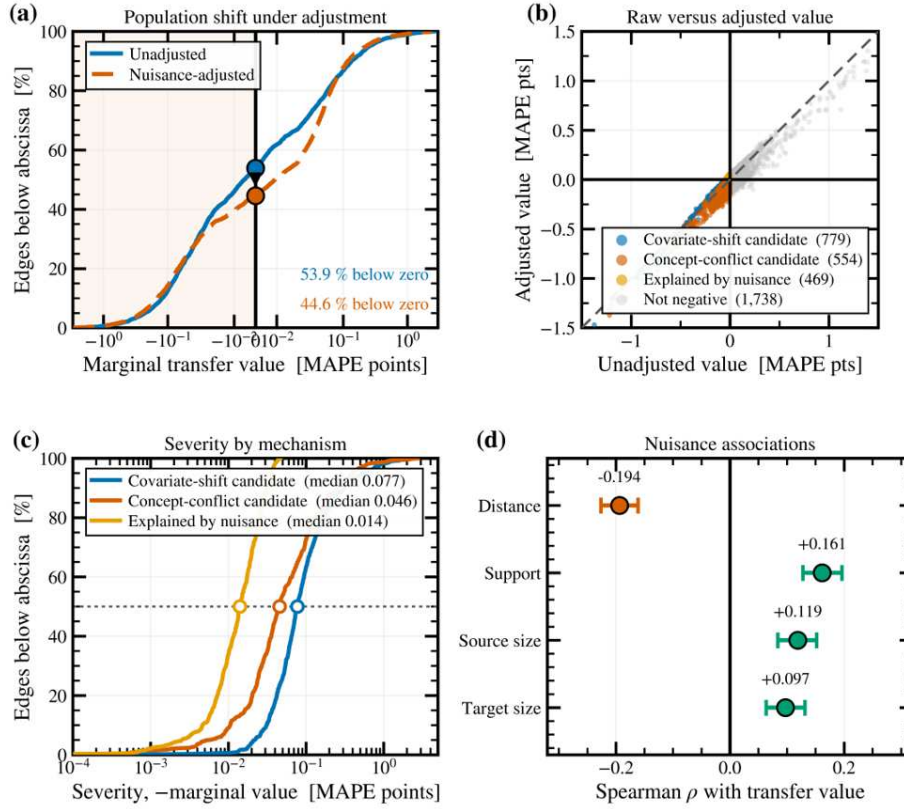


Fig. 11 | Measured factors explain only part of the harmful transfer. (a) Harmful-edge classification after adjustment for campaign size, distance, and support; (b) raw versus adjusted transfer value; (c) severity of the adjusted candidate classes; and (d) associations with the measured nuisance variables. Adjustment mainly removes weak conflicts, while a stronger negative-transfer tail remains unresolved.

3.4.3 Coalition value is real but transport-game dependent

Removal from the complete source pool represents a single marginal contribution and can hide redundancy. Permutation Shapley valuation makes coalition dependence explicit (Fig. 12a). Forty-five of 60 campaigns have positive mean coalition value as sources; the mean Monte-Carlo standard error is 0.248 MAPE points (Fig. 12b). Removal marginal and Shapley value remain strongly rank-correlated ($\rho=+0.781$), but their signs agree in only 76.6% of comparisons (Fig. 12c). A campaign can therefore be redundant in the complete archive yet valuable in a smaller coalition, or vice versa.

The more consequential sensitivity is to the transport game itself. Source rankings from Gaussian-support and inverse-distance games agree strongly (Spearman $\rho=+0.833$), whereas Gaussian support versus the truncated ten-nearest-campaign game gives only $\rho=+0.060$; inverse distance versus the truncated game is likewise weak at $\rho=+0.198$. The number of mean-positive sources is comparatively stable (45, 45, and 41), but the exact ordering is not. Continuous weighting rules, therefore, recover a similar coalition hierarchy, whereas truncating the source set poses a materially different valuation problem.

This distinction bounds how campaign value may be used. Precise Shapley ranks must be reported with the game that generated them and should not be interpreted as an intrinsic property of an experiment. By contrast, the majority sign of the directed removal graph is reproduced for 94.3% of edges across all three games. The prospective replication priorities, therefore, rely on directed transfer structure and support geometry rather than on a single Shapley ordering.

The efficiency axiom closes to a maximum absolute residual of 3.375×10^{-14} MAPE points, far inside the predeclared 10^{-8} gate (Fig. 12d). This establishes numerical accounting, not learner invariance. TRACE

provides an estimator-independent test of campaign heterogeneity, but it does not recompute the same Shapley game; neural-network, Gaussian-process, or other learner-specific campaign rankings remain open robustness questions.

Fig. 10, Fig. 11 and Fig. 12, therefore establish the second central result: historical CHF evidence is neither uniformly beneficial nor exchangeable across experiments. Its usefulness depends on direction, thermohydraulic support, the surrounding coalition, and, in the case of precise Shapley ordering, the declared transport rule. Coalition-dependent valuation of historical CHF campaigns is presented in Fig. 12.

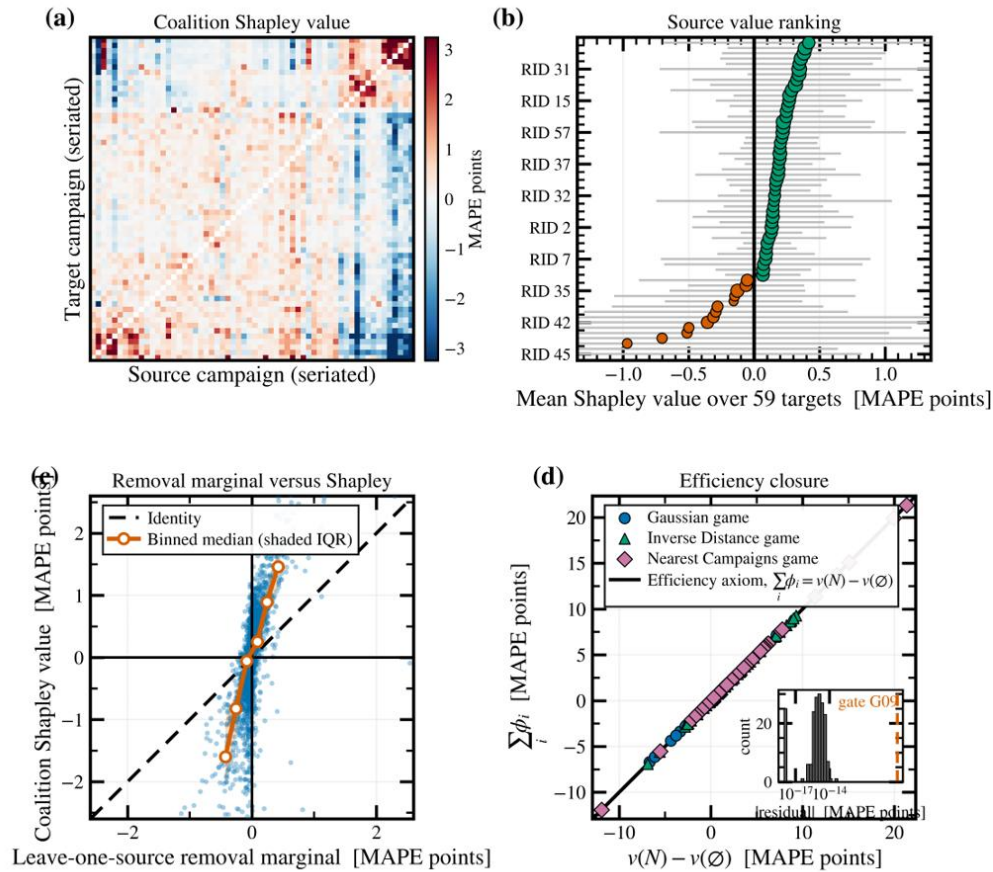


Fig. 12 | Coalition-dependent value of historical CHF campaigns. (a) Target-by-source permutation-Shapley map; (b) source ranking by mean coalition value; (c) removal marginal versus Shapley value; and (d) numerical closure of the Shapley efficiency relation. Broad usefulness is reproducible, but precise campaign ranking depends on the declared transport game.

Table 10 | Directed campaign transfer, reciprocal asymmetry, nuisance adjustment, and coalition-value diagnostics over all 3,540 ordered source–target pairs.

Statistic	Definition or evaluation set	Observed result	Permitted interpretation
Helpful edges	Positive game-mean removal marginal; majority sign	1,537 (43.4%)	Source improves the declared game
Harmful edges	Negative game-mean removal marginal; majority sign	1,802 (50.9%)	Negative transfer in the declared game
Game-dependent edges	No majority-supported sign	201 (5.7%)	Sign-sensitive evidence
Unanimous sign	Same sign in all three games	806 (22.8%)	Strongest sign-reproducibility class
The majority sign the agreement	At least two of three games	3,339 (94.3%)	Passes G11

Statistic	Definition or evaluation set	Observed result	Permitted interpretation
Median edge magnitude	Absolute game-mean removal value	0.042 MAPE points	Most individual effects are small
Distance contrast	Closest vs most distant strata	31% vs 64% harmful	Harm increases with separation
Support contrast	Lowest vs highest support strata	68% vs 27% harmful	Harm decreases with source support
Reciprocal sign discordance	1,770 unordered campaign pairs	30.0% discordant; 70.0% concordant	Value is directional
Reciprocal rank association	Vrem(s→t) vs Vrem(t→s)	Spearman $\rho=+0.565$	Only moderate reciprocal symmetry
Nuisance-explained negatives	Harmful edge becomes non-negative after ridge adjustment	469 of 1,802 (26%)	The measured nuisance explains part of the weak tail
All sign flips after adjustment	All directed edges	550	Adjustment mainly reclassifies near-zero edges
Mean-positive Shapley sources	Mean over other targets	45 of 60	Coalitional usefulness is uneven
Removal–Shapley association	Across directed source–target values	Spearman $\rho=+0.781$; sign agreement 76.6%	Related, non-identical value concepts
Gaussian vs inverse-distance rank	Game-specific source Shapley ranking	$\rho=+0.833$	Continuous games agree strongly
Gaussian vs 10-nearest rank	Game-specific source Shapley ranking	$\rho=+0.060$	Truncation changes ordering
Inverse-distance vs 10-nearest rank	Game-specific source Shapley ranking	$\rho=+0.198$	Truncation changes ordering
Mean-positive sources by game	Gaussian / inverse-distance / 10-nearest	45 / 45 / 41	Sign/count is more stable than rank
Shapley efficiency residual	Maximum absolute closure error	3.375×10^{-14} MAPE points	Numerical audit passes G09

3.5 Sparse labels recover campaign state; acquisition-policy optimality remains open

Before any target CHF is revealed, outcome-blind transport diagnostics still contain useful information about likely difficulty. Source disagreement, helpful-prior weight, and related quantities show structured associations with realized error (Fig. 13a–c). Used only as a ranking variable, nearest-source distance identifies a safer 50% subset with a MAPE of 2.70%, compared with the 4.52% zero-shot record-weighted base rate (Fig. 13d). These diagnostics can support cautious commissioning. Still, predictive association is not evidence that a complex query score is optimal. Outcome-blind transport diagnostics and their association with realized prediction risk are evaluated in Fig. 13.

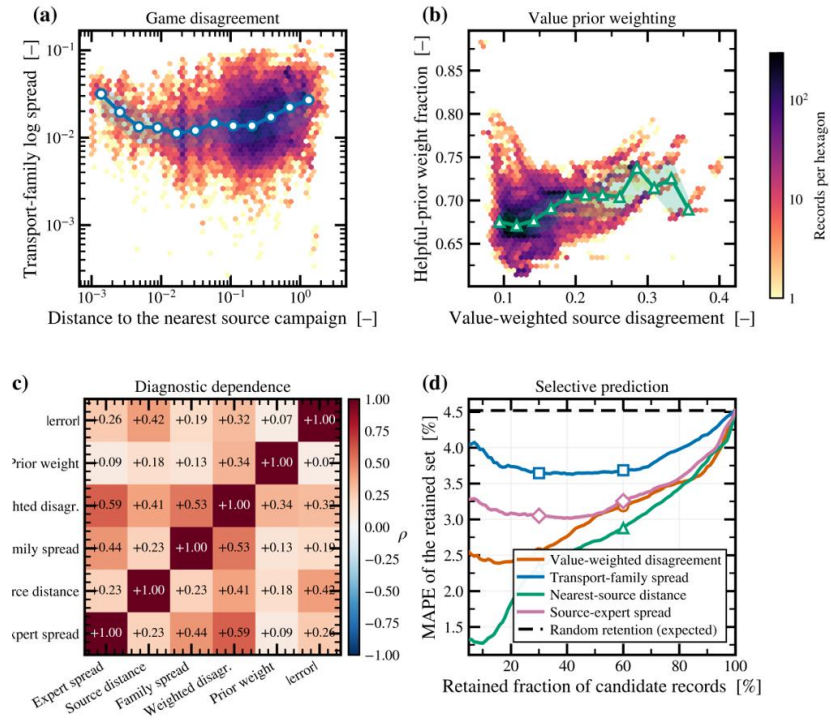


Fig. 13 | Outcome-blind transport diagnostics and realized prediction risk. (a) Transport-family spread versus nearest-source distance; (b) helpful-prior weight versus value-weighted source disagreement; (c) dependence among the five diagnostics and realized absolute error; and (d) selective-prediction risk–coverage curves. Nearest-source distance is the strongest single warning variable in this audit.

3.5.1 Sparse measurements recover a predominantly low-dimensional campaign state

The adaptation test examines whether the revealed target measurements correct the unresolved campaign residual. With no target labels, campaign-macro MAPE is 7.474%. After five TWIN-VALUE-selected measurements are revealed and assimilated through the regularized Bayesian update, it falls to 5.650%, a 24.4% reduction (Fig. 14a; Table 11). The paired campaign-bootstrap lower bound for the gain is +0.515 MAPE points. G07A and G07B, therefore, support sparse adaptation, while G07C records the deliberate 25% stretch target as narrowly failed.

The gain is not specific to the value-guided policy: with the same update, repeated random selection reaches 5.599% macro MAPE at $k=5$, maximin coverage reaches 5.620%, whereas uncertainty-driven acquisition remains at 7.223%. This pattern is consistent with a low-dimensional campaign correction: when the dominant recoverable information is an experiment-wide level offset, several reasonably dispersed measurement sets identify it similarly, while repeatedly selecting the hardest or most atypical conditions can be unrepresentative of that offset.

Adaptation is nevertheless heterogeneous. At $k=5$, 37 of 58 eligible campaigns improve and 21 degrade; mean gain is +1.51 but median gain only +0.27 MAPE points (Fig. 14c). The update rule is also load-bearing: under uncertainty-driven sampling, an unregularized offset reaches 50.4% macro MAPE at $k=10$, whereas the regularized Bayesian update remains controlled (Fig. 14b). These results support residual monitoring and rollback rather than unconditional assimilation. Sparse adaptation of previously unseen campaigns is quantified in Fig. 14.

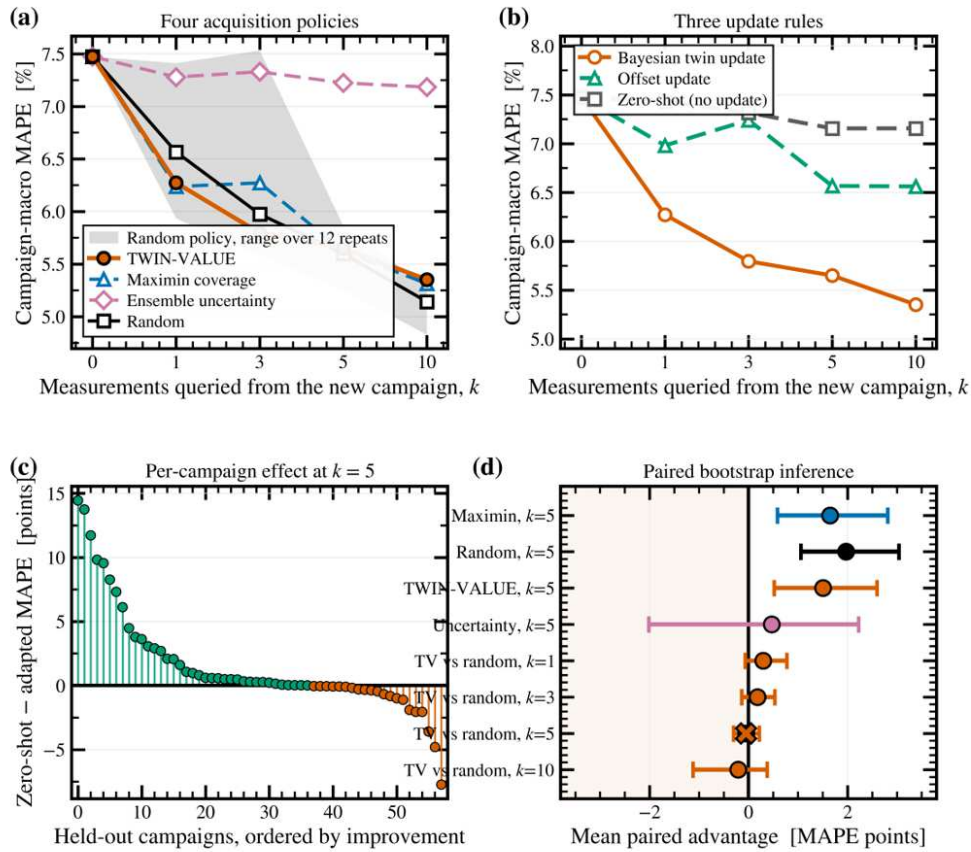


Fig. 14 | Sparse adaptation of previously unseen campaigns. (a) Campaign-macro error as measurements are revealed under four acquisition policies; (b) regularized versus unregularized updating; (c) campaign-wise five-measurement gains; and (d) paired comparison with repeated random sampling. Sparse updating is beneficial, but the more complex acquisition policy does not show statistically significant superiority.

3.5.2 The proposed acquisition policy does not outperform random selection

The success of sparse assimilation does not validate the acquisition strategy. Using a consistent advantage sign convention of random minus TWIN-VALUE, the point differences are +0.293, +0.179, -0.051, and -0.210 MAPE points at $k=1,3,5,10$, with 95% intervals $[-0.070, +0.774]$, $[-0.138, +0.528]$, $[-0.305, +0.216]$, and $[-1.123, +0.374]$, respectively. Every interval contains zero, and the point estimate changes sign as the budget changes. G08B therefore fails at $k=5$, and no tested budget supports a claim of superiority.

The $k=5$ point estimate remains inside the predeclared 0.25-point non-inferiority tolerance (TWIN-VALUE minus random = +0.051), but that is not evidence that the additional scoring complexity is useful. The supported algorithmic result is the regularized campaign-state update. Measurement-policy design remains open, with simple random or maximin coverage competitive in the present replay and uncertainty-driven selection distinctly poor under sparse budgets.

Table 11 | Offline sparse-assimilation and acquisition-policy decision criteria, with policy comparisons separated from update benefit.

Estimand	Comparison	Observed result	Decision
Zero-shot macro MAPE	Before target labels	7.474%	Reference state
Five-shot macro MAPE	TWIN-VALUE + regularized Bayesian update	5.650%	Adapted state
Five-shot macro MAPE	Repeated random + same update	5.599%	Statistically indistinguishable from TWIN-VALUE
Five-shot macro MAPE	Maximin coverage + same update	5.620%	Statistically indistinguishable from TWIN-VALUE
Five-shot macro MAPE	Uncertainty acquisition + same update	7.223%	Poor sparse-budget policy

Estimand	Comparison	Observed result	Decision
Relative adaptation gain	Zero-shot vs TWIN-VALUE k=5	24.4%	G07A pass; G07C stretch fail
Bootstrap lower bound	Paired campaign gain, k=5	+0.515 MAPE points	G07B pass
Policy advantage, k=1	Random – TWIN-VALUE	+0.293; 95% CI [−0.070,+0.774]	Contains zero
Policy advantage, k=3	Random – TWIN-VALUE	+0.179; 95% CI [−0.138,+0.528]	Contains zero
Policy advantage, k=5	Random – TWIN-VALUE	−0.051; 95% CI [−0.305,+0.216]	G08B fail
Policy advantage, k=10	Random – TWIN-VALUE	−0.210; 95% CI [−1.123,+0.374]	Contains zero
Non-inferiority tolerance	TWIN-VALUE – random, k=5	+0.051 points	G08A pass; margin 0.25
Campaign response	Improved/degraded at k=5	37 / 21 of 58	Heterogeneous adaptation
Per-campaign gain	Mean / median at k=5	+1.51 / +0.27 points	Mean driven by difficult campaigns
Unregularized instability	Offset update; uncertainty policy; k=10	50.4% macro MAPE	Regularization is essential

3.6 Physical closure and prospective falsification define the boundary of the surrogate

The adapted prediction is finally coupled to the deliberately limited thermal calculation in Eq. (14). Figure 15a compares the numerically solved enthalpy rise with the analytical energy balance; the maximum relative deviation is 7.1×10^{-14} . Representative equilibrium-quality trajectories supply only axial energy accounting (Fig. 15b). Across 59 conditions, the adapted versus zero-shot predicted limits have a median change of +0.2% and a range from −26.0% to +76.0% (Fig. 15c). Across 238 evaluated zero-shot and adapted states, the independently archived maximum relative conservation residual is 3.073×10^{-12} , about 33 times below the predeclared 10−10 threshold (Fig. 15d).

These calculations verify arithmetic conservation and computational coupling; they do not convert the CHF surrogate into a multiphase solver or digital twin. Pressure drop, wall temperature, void fraction, turbulence, interfacial transport, transient state, margin-to-CHF, and the physical location or mechanism of the boiling crisis remain outside the model. The one-dimensional balance is a bounded physics interface and software-verification check, not evidence of reactor-level fidelity. Verification of the bounded one-dimensional thermal interface is provided in Fig. 15.

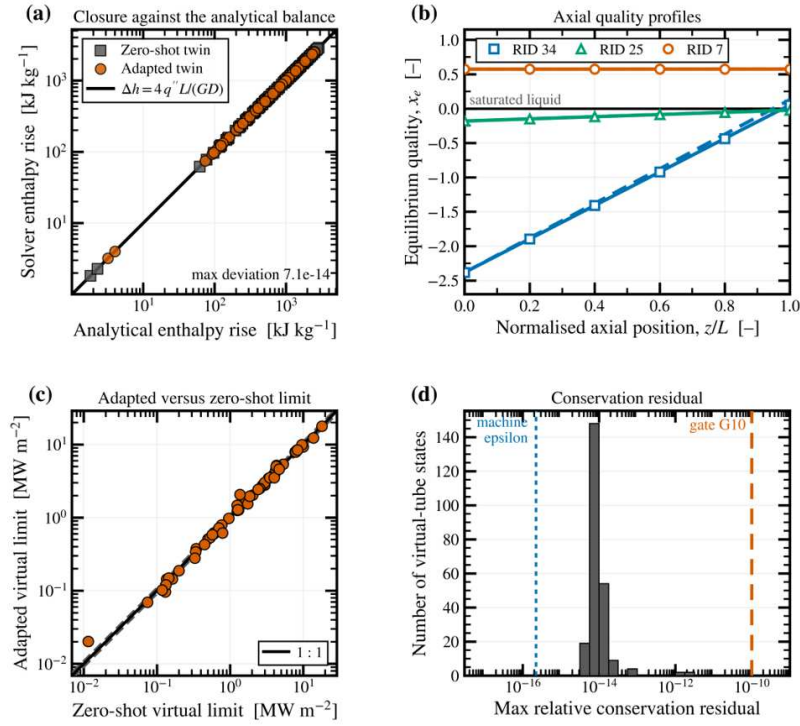


Fig. 15 | Verification of the bounded one-dimensional thermal interface. (a) Numerical enthalpy rise compared with the analytical energy balance; (b) representative axial equilibrium-quality profiles; (c) adapted versus zero-shot predicted limits; and (d) conservation residuals across evaluated states. The calculation verifies energy accounting only and does not establish multiphase-flow or reactor-scale fidelity.

The framework closes with a proposed prospective test rather than another retrospective fit. Thirty outcome-blind candidate conditions lie within the documented operating and geometry envelope, span 16 archived campaign envelopes, and carry frozen predicted CHF values from 0.011 to 10.99 MW m^{-2} with 95% intervals (Fig. 16a–c). Priority rank is not significantly monotone with relative interval width (Spearman $\rho = -0.196$, $p = 0.298$; Fig. 16d), consistent with ranking that combines support, disagreement, and transfer risk rather than simply selecting the widest intervals.

These conditions define a falsification matrix for the central cross-campaign claims. A valid future experiment would compare the frozen zero-shot predictions and intervals with newly measured CHF before any refitting, then repeat the predeclared sparse-update comparison. The present study does not claim prospective success, external preregistration, or independent replication; the value of the matrix is that a future result can contradict the current model rather than being absorbed retrospectively. The outcome-blind matrix for future experimental falsification is presented in Fig. 16.

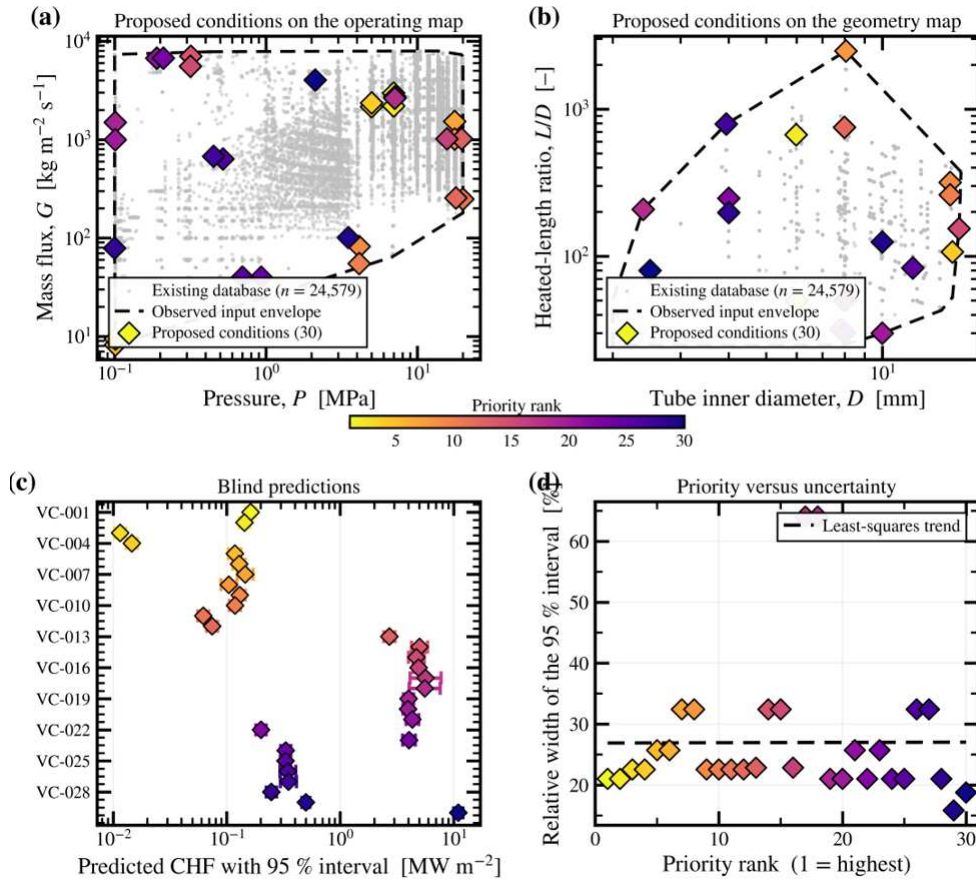


Fig. 16 | Outcome-blind matrix for future experimental falsification. (a) Proposed conditions in pressure–mass-flux space; (b) corresponding geometry coverage; (c) frozen CHF predictions with 95% intervals; and (d) priority rank versus relative interval width. The matrix is designed to test zero-shot transfer, uncertainty, and sparse adaptation using genuinely new measurements.

3.7 Independent hierarchical analysis triangulates campaign provenance and response topology

Every result above is conditional on the primary physics-anchored tree-ensemble framework. TRACE therefore asks a deliberately different question with a different estimator and target on the same locked archive. It is independent of the estimator, random-effects representation, and response-field topology, but not of the underlying data. Its role is triangulation: to test whether campaign-level structure and more than one reproducible response pattern remain visible without using the primary champion or its transfer utility.

3.7.1 Campaign effects are small, predominantly level-like, and reproducible

After averaging each RID/seed across its fold-level estimates and summarizing the 60 campaigns, the TRACE campaign intercept has a sample standard deviation of $1.751\text{e-}03$, a range of $-6.345\text{e-}03$ to $+7.082\text{e-}03$, and a median absolute magnitude of $6.729\text{e-}04$. The median absolute magnitude across the five campaign-slope coefficients is $1.787\text{e-}04$; under this explicit aggregation, the intercept is $3.76\times$ larger. Intercept ranks are reproduced across the three seed pairs, with Spearman ρ values of 0.982, 0.990, and 0.981. Thus, the hierarchical estimator independently supports a stable, predominantly level-like provenance state, the same low-dimensional structure that motivates stronger shrinkage of slopes in the sparse Bayesian update, while the absolute effect size remains small.

That agreement should not be overstated. Across the 36-fold seed selections, the weakest random-effect penalty (0.3) is selected 19 times, the strongest (10) 16 times, and the intermediate value 1.0 once. Averaged over the complete regularization grid, validation campaign-macro MAPE changes by only 0.0096 points across random-effect penalties. The fitted campaign effects are therefore reproducible, but their regularization strength is only weakly identified by predictive validation; campaign structure is present but not the dominant source of TRACE predictive accuracy.

3.7.2 A two-pattern response topology survives the predeclared stability rule

Only $K=2$ satisfies all six predeclared topology criteria. Relative to $K=1$, it improves BIC by 5,226.2, has a silhouette of 0.270, a median campaign-bootstrap ARI of 0.785, an RBF-versus-local ARI of 0.285, a minimum regime weight of 0.478, and support in at least 43 campaigns (Table 12). $K=3-5$ gain further BIC, as more

flexible nested descriptions can, but fail one or more stability, cross-estimator, or occupancy criteria. The decision is therefore based on stability rather than BIC alone.

Table 12 | TRACE response-topology selection under the predeclared multi-criterion stability rule.

K	ΔBIC vs K=1	Silhouette	Median bootstrap ARI	RBF/local ARI	Min. weight	Min. campaigns	Decision
1	0.0	,	1.000	1.000	1.000	58	Fail
2	5,226.2	0.270	0.785	0.285	0.478	43	Pass; selected
3	7,815.1	0.261	0.679	0.241	0.152	39	Fail
4	8,659.5	0.176	0.490	0.201	0.210	28	Fail
5	9,890.6	0.211	0.538	0.140	0.083	10	Fail

R1 contains 8,732 records in 57 campaigns, and R2 contains 15,847 records in 45 campaigns (Table 13). Their fitted elasticities differ most strongly in mass flux, L/D, and normalized inlet subcooling. Crucially, 42 of 60 campaigns contain both response labels, and only 18 contain one: response pattern and experimental provenance are partially crossed rather than nested. A harmful campaign-to-campaign edge, therefore, cannot be explained simply by assigning each experiment to a single regime.

Table 13 | TRACE phenomenological response-pattern signatures. Elasticities are derivatives of the fitted response field, not measured mechanistic sensitivities.

Pattern	Records	Campaigns	e_Pr	e_G	e_D	e_L/D	e_Nsub
R1	8,732	57	-0.112	-0.508	-0.292	-0.606	+1.417
R2	15,847	45	-0.085	-0.306	-0.146	-0.799	+0.984

3.7.3 Estimator independence strengthens the heterogeneity claim, not the accuracy claim

As a predictive baseline, TRACE_SHARED_FIELD yields a campaign-macro MAPE of 12.112% (95% campaign-bootstrap interval: 9.879–14.938%), better than its own internal Bowring comparison at 16.448% but substantially weaker than the primary validation-selected champion at 7.474% (Table 14). LOCAL_LINEAR_AUDIT gives 13.203%. The TRACE Bowring value belongs only to the separately archived TRACE pipeline and is not substituted for the primary Bowring benchmark of 16.084% in Table 6. The scientific value of TRACE is therefore not to improve the headline error; it is to show that reproducible provenance effects and a stable two-pattern topology appear under a materially different estimator.

Together, the random-effect and topology audits refine the latent-state interpretation. Provenance behaves predominantly as a small level-like offset, while operating conditions traverse at least two stable phenomenological response patterns. Because most campaigns contain both patterns, neither source replaces the other: campaign state and response topology are partially crossed dimensions of heterogeneity. The archive still lacks the surface, wall-temperature, film, hydraulic-development, and metrology observations required to convert either dimension into a mechanism label.

Table 14 | Predictive performance internal to the independent TRACE audit. Values are not used to select or revise the primary campaign-held-out champion.

Method	Campaign-macro MAPE	95% CI	Macro bias	Macro RMSPE	Within $\pm 20\%$	k95
TRACE shared field	12.112%	9.879–14.938%	+0.654%	15.014%	83.0%	1.313
Local-linear audit	13.203%	9.723–17.551%	+1.133%	17.567%	85.0%	1.394
Bowring (TRACE pipeline)	16.448%	13.211–19.948%	-6.759%	20.790%	70.7%	1.540

4. Discussion

The principal finding is not simply that machine learning lowers CHF error. It is that CHF generalization changes when the experimental campaign is treated as the unit of evidence. The archive is strongly concentrated, and whole-campaign testing exposes a difficulty structure that persists even after the model

family, validation-only selection, and physics anchoring are varied. This makes cross-experiment prediction both an evidence-transport problem and a regression problem.

The concentration decomposition sharpens that claim quantitatively. Capping each campaign at 500 records changes the frozen champion's weighted MAPE from 4.517% to 5.676%, recovering 39.2% of the distance to the 7.474% equal-campaign value. The remaining 60.8% is not a thermophysical percentage; it is residual campaign-difficulty heterogeneity after a weight-only perturbation. The largest tercile is simultaneously the most heavily sampled and the easiest on equal-campaign MAPE, so pooled benchmarks are favored both by row count and by which experiments generated those rows. For concentrated scientific archives, record-weighted and experiment-weighted metrics should therefore be reported together.

Physics anchoring still delivers a substantial predictive benefit. The Bowring-residual ensemble improves average, tail, and campaign-level error while preserving an auditable engineering reference. What the anchor does not do is erase experimental provenance. Residual extremes and hard campaigns identify where the fixed closure and recorded coordinates transfer poorly; without wall-temperature histories, surface descriptors, film measurements, or flow-regime observations, those residuals remain targets for mechanism-resolving experiments rather than evidence for a named boiling-crisis mechanism.

The transfer graph adds a structural result that pooled prediction cannot reveal. Slightly more than half of the majority-classified directed edges are harmful, and harm increases with thermohydraulic distance. It decreases with source support, and 26% of harmful edges become non-negative after adjustment for measured nuisance covariates. More importantly, transfer is directional: 30.0% of reciprocal campaign pairs disagree in sign, and the reciprocal-rank association is only +0.565. Historical experiments, therefore, do not partition into mutually compatible and incompatible groups. A broad experiment can support a narrow target without the narrow experiment returning equal value, so any evidence-selection rule must respect source→target direction.

Coalition value is even more conditional. Removal and Shapley values are related but not interchangeable, and the exact Shapley source ordering depends strongly on how transport support is defined. Continuous Gaussian and inverse-distance games agree at $\rho=+0.833$, whereas the truncated nearest-campaign game is nearly unrelated to Gaussian ranking ($\rho=+0.060$). This comparison identifies the more robust quantity as the directed sign structure, with 94.3% majority agreement across games, rather than a universal scalar ranking across experiments. Shapley scores should be reported with their learner, utility, and transport game.

TRACE provides independent triangulation of the missing-state interpretation. Under a transparent RID/seed aggregation, campaign intercepts are about 3.76 times larger in median absolute magnitude than campaign slopes and reproduce across seeds at $\rho\approx 0.981-0.990$. At the same time, only $K=2$ satisfies the full topology-stability rule, yet 42 campaigns contain both response patterns. Provenance and response topology are therefore partially crossed: a small level-like facility state can coexist with condition-dependent response structure inside the same experiment. The random-effect regularization is weakly identified, and TRACE shares the same archive, which supports estimator independence rather than causal or external validation.

Uncertainty follows the same evidentiary unit. A nominal 95% pooled interval achieves 91.85% record-weighted coverage, but only 82.49% campaign-macro coverage and 34.92% lower-decile coverage. Campaign-robust calibration raises the latter two to 92.17% and 76.99%, but requires wider intervals. This trade-off shows that protecting a typical held-out experiment is more demanding than covering records dominated by a few large campaigns. Even the robust result remains empirical, and evidence distance plus model disagreement should remain warning signals for a genuinely new facility.

The few-shot replay identifies the most actionable component. Five revealed measurements reduce macro MAPE from 7.474% to 5.650% with a positive paired lower confidence bound, while random and maximin selection achieve essentially the same five-shot result, and uncertainty-driven selection performs poorly. Across $k=1, 3, 5, 10$, every TWIN-VALUE-versus-random interval contains zero, and the point difference changes sign. The supported contribution is therefore regularized state recovery, not sophisticated active learning. The TRACE intercept result provides an independent explanation: with only a few labels, much of the recoverable information is a campaign-wide-level correction, so several dispersed samples can identify it, while hard-point chasing may be counterproductive.

Regularization is essential to that interpretation. The unregularized offset can react catastrophically to an extreme early observation, reaching 50.4% macro MAPE at $k=10$ under uncertainty-driven sampling, whereas the shrinkage update remains controlled. Yet the exact prior precisions, observation-noise scale, and clipping limits are implementation choices, not physical constants, and a complete test-independent sensitivity surface is not available. A practical commissioning workflow should monitor residuals after every reveal and retain or revert to the zero-shot prediction when adaptation degrades the target campaign.

The one-dimensional thermal calculation is deliberately narrower than the data-driven framework around it. Its conservation residual verifies that the predicted limiting flux is propagated consistently through the specified steady enthalpy balance; it establishes nothing about pressure drop, void fraction, wall temperature, interfacial structure, CHF location, or transient response. Calling the framework an offline adaptive campaign-initialized surrogate rather than a digital twin is therefore part of the scientific result: the terminology matches the actual physics and information loop.

Prospective falsification remains the decisive external test. The 30-condition matrix freezes query coordinates, predictions, intervals, and priority before any new CHF measurement is made. Its rank is not merely an uncertainty ranking; it can test zero-shot transfer, empirical coverage, and sparse updating in the presence of genuinely new evidence. Until those measurements are obtained independently and reported without retrospective refitting, the present study supports a strong hierarchy of retrospective claims but not of prospective validation, safety certification, causal facility diagnosis, or operational-twin status.

5. Limitations

The results apply only to the archived water-CHF data for uniformly heated vertical round tubes. They should not be generalized directly to other fluids, geometries, heating conditions, or new facilities. Campaign identity is an experimental provenance label, not a physical mechanism; the present data cannot separate the effects of surface condition, instrumentation, facility design, omitted variables, and true CHF-response differences. Likewise, transfer values and Shapley rankings depend on the chosen model and transport definition, so they describe predictive usefulness rather than intrinsic experimental value.

The uncertainty and adaptation results are also retrospective. Campaign-aware calibration improves coverage for the archived experiments, but it does not guarantee the same coverage for an arbitrary future facility. Sparse Bayesian updating improves prediction, yet its regularization settings are not universal, and the TWIN-VALUE acquisition policy does not statistically outperform repeated random selection. TRACE provides an independent estimator check, but it uses the same database and therefore is not an independent experimental validation. Finally, the one-dimensional thermal calculation verifies energy conservation only; it does not model two-phase flow, wall temperature, void fraction, pressure drop, CHF location, or transient behavior. The proposed 30-condition matrix remains a future falsification test, not a completed prospective validation.

6. Conclusions

This study tests whether historical CHF campaigns behave as exchangeable evidence once the measured operating and geometric variables are supplied. They do not. Treating the campaign as the common unit of validation, uncertainty, data value, and adaptation reveals a consistent experiment-level structure that pooled record analysis cannot resolve.

1. Whole-campaign testing exposes a persistent generalization penalty. The validation-selected physics-anchored surrogate reduces record-weighted MAPE from 9.670% to 4.517% and campaign-macro MAPE from 16.084% to 7.474%, improving 52 of 60 unseen campaigns. Across all seven predictors, campaign-macro error remains 1.65–1.69 times the record-weighted value. A weight-only 500-record cap closes 39.2% of the champion's record-to-campaign gap, leaving 60.8% as residual campaign-difficulty heterogeneity. Historical sampling concentration matters, but it does not explain the full cross-experiment penalty.

2. Historical experiments are non-exchangeable, and their value is directional. Of 3,540 directed source-to-target transfers, 1,802 (50.9%) are harmful, 1,537 (43.4%) are helpful, and 201 (5.7%) have no majority-

supported sign. Harmful transfer rises from 31% to 64% across increasing thermohydraulic-distance strata and falls from 68% to 27% as relative source support improves. Among the 1,770 reciprocal campaign pairs, 30.0% reverse the sign, and the two directions have only a moderate rank association (Spearman $\rho=0.565$). Experimental usefulness is therefore a property of the ordered source-target relation, not a symmetric compatibility label.

3. The transfer sign is more robust than a precise coalition ranking. At least two of the three transport games agree on the sign for 3,339 of 3,540 directed edges (94.3%). Removal marginal and Shapley value are related ($\rho=0.781$) but agree in sign for only 76.6% of comparisons. Shapley source rankings agree strongly between Gaussian-support and inverse-distance games ($\rho=0.833$), but weakly between Gaussian-support and the truncated 10-nearest-campaign game ($\rho=0.060$). Directed sign structure is therefore the more defensible basis for replication priorities; exact Shapley ranks remain conditional on the declared game and learner.

4. Campaign-level uncertainty and sparse adaptation require explicit treatment of provenance. At nominal 95%, pooled calibration gives 82.49% campaign-macro coverage and 34.92% lower-decile coverage, whereas the campaign-aware rule raises these to 92.17% and 76.99%, with a median within-campaign width inflation of $1.47\times$. Five revealed measurements reduce unseen-campaign macro MAPE from 7.474% to 5.650%, a 24.4% gain with a paired bootstrap lower bound of $+0.515$ MAPE points. TWIN-VALUE does not beat repeated random acquisition with statistical confidence at any tested budget; the supported result is regularized state recovery, not acquisition-policy superiority.

5. The remaining campaign structure is reproducible but is not a uniquely identified CHF mechanism. In the independent TRACE analysis, campaign-intercept ranks reproduce across seeds at $\rho=0.981\text{--}0.990$, and the median absolute intercept is $3.76\times$ the median campaign-slope magnitude, supporting a predominantly level-like residual state. A two-pattern topology is selected by the predeclared stability rule ($\Delta\text{BIC}=5,226.2$, silhouette=0.270, median bootstrap ARI=0.785), yet 42 of 60 campaigns contain both patterns, so response pattern and provenance are not interchangeable. The thermal interface independently closes the imposed energy balance with a maximum archived conservation residual of 3.073×10^{-12} , below the 10–10 gate; this verifies conservation only, not pressure drop, void fraction, wall temperature, CHF location, or multiphase fidelity.

Taken together, the results support a campaign-aware, physics-anchored adaptive CHF surrogate in which prediction, uncertainty, experimental value, and sparse updating are evaluated within a single consistent evidentiary unit. The practical implication is not simply to collect more CHF records. New experiments should deliberately fill weakly supported operating regions, repeat conditions associated with strong directional disagreement, and report surface, heater, hydraulic, and metrology information that can turn the present provenance signal into a physically interpretable state. The frozen 30-condition matrix provides the next decisive test: genuinely new measurements must challenge the predictions before any refitting is allowed.

Nomenclature

Symbol/abbreviation	Definition	Unit/status
Thermophysical and geometric quantities		
P	Pressure	MPa
P_{crit} ; G_0 , D_0 , L_0 , h_0 , q_0	Critical pressure and fixed reference scales are used for normalization	22.064 MPa; 1000 kg m ⁻² s ⁻¹ ; 0.01 m; 1 m; 10 ⁶ J kg ⁻¹ ; 1 W m ⁻²
G	Mass flux	kg m ⁻² s ⁻¹
D	Tube inner diameter	m
L	Heated length	m
z	Axial coordinate	m
h	Specific enthalpy	J kg ⁻¹
$h_{f, \text{sat}}$	Saturated-liquid specific enthalpy	J kg ⁻¹
h_{fg}	Latent enthalpy of vaporization	J kg ⁻¹
$\Delta h_{\text{sub}, \text{in}}$	Inlet subcooling enthalpy deficit	J kg ⁻¹
q_{CHF}	Measured critical heat flux	W m ⁻²
q_B	Critical heat flux predicted by the Bowring anchor	W m ⁻²
$\hat{q}$	Model-predicted critical heat flux	W m ⁻²

Symbol/abbreviation	Definition	Unit/status
q''	Uniform imposed wall heat flux	W m^{-2}
ρ_l, ρ_g	Saturated liquid and vapor densities	kg m^{-3}
μ_l	Saturated-liquid dynamic viscosity	Pa s
σ	Surface tension	N m^{-1}
T_{sat}	Saturation temperature	K
x_e	Thermodynamic equilibrium quality	dimensionless
$A, C, f_1\text{--}f_4, n$	Bowring correlation terms and pressure functions	Eq. (1)
Prediction, calibration, and evaluation		
$\mathbf{u}_i, \tilde{\mathbf{u}}_i$	Dimensionless pre-CHF feature vector for record i and its outer-training-standardized form	dimensionless
$\eta_i^{(m)}, \hat{\eta}_i^{(m)}$	Model-specific logarithmic learning target and its prediction for candidate model m	dimensionless
$q_{\text{ref},i}^{(m)}, \hat{q}_i^{(m)}$	Model-specific reference heat flux and reconstructed candidate CHF prediction	W m^{-2}
$h_{\theta}^{(m)}, \mathcal{M}_{\text{res}}, \mathcal{M}_{\text{dir}}$	Candidate-model function, candidate set, and residual/direct model subsets	model quantities/sets
w_m	Non-negative convex weight of candidate model m	dimensionless
e_{ij}	Absolute percentage error of record i in campaign j	%
MAPE_{rw}	Record-weighted mean absolute percentage error	%
$\text{MAPE}_{\text{macro}}$	Mean of campaign-level mean absolute percentage errors.	%
$N, J, n_j, N_{\text{train}}, J_{\text{train}}$	Record, campaign, per-campaign, outer-training-record, and outer-training-campaign counts	counts
α	Miscoverage probability	dimensionless
a_i	Absolute log-residual nonconformity score	dimensionless
c_α	Pooled conformal log-scale half-width	dimensionless
$c_{\text{CR},\alpha}$	Campaign-robust conformal log-scale half-width	dimensionless
$I_\alpha, I_{\text{CR},\alpha}$	Pooled and campaign-robust prediction intervals	W m^{-2}
Campaign transport and adaptive updating		
$\mathbf{x}; S, s, t$	Generic query condition; source set, source campaign, and target campaign	query/indices/set
$\xi, \xi_i, K, \ell_s, \mathcal{N}_{12}^{(s)}(\xi)$	Target/source transport coordinates, kernel, source length scale, and 12-nearest-record set	transport-space quantities
$\mathcal{G}, g, \omega_s^{(g)}$	Transport-game set, game index, and between-source weight	set/index/dimensionless
$V_{\text{rem}}^{(g)}(s \rightarrow t)$	Game-specific directed source-removal value for target t	MAPE points
$\phi_{s,t}^{(g)}, U^{(g)}(S; t)$	Game-specific permutation Shapley value and predictive utility	MAPE points / negative MAPE
$V_{\text{mean},s,t}, \beta_{s,t}, \mathbf{c}_{s,t}$	Game-mean removal value, ridge coefficients, residual, and standardized nuisance-covariate vector	model quantities
$V_{\text{adj},s,t}$	Nuisance-adjusted directed transfer value	MAPE points
$\mathbf{B}_k, \mathbf{b}(\mathbf{x}), \delta_k$	Few-shot design matrix, update-basis vector, and revealed zero-shot log-residual vector	dimensionless
Λ_0, Σ_k	Prior precision and posterior covariance matrices	model quantities
μ_k, σ_δ^2	Posterior coefficient mean and Bayesian update observation variance	model quantities
A_{TV}	Outcome-blind TWIN-VALUE acquisition score; retained as a policy/workflow label, not a digital-twin claim	dimensionless

Symbol/abbreviation	Definition	Unit/status
D_{div} , D_{dom} , ΔV_{post} , S_{src} , U_{ens}	Diversity, domain, posterior-value, source-support, and ensemble-disagreement scores	dimensionless
i , j , k , m , g	Record, campaign, revealed-measurement, candidate-model, and transport-game indices	indices
Abbreviations		
CHF	Critical heat flux	
RID	Reference-identifier campaign	
MAPE	Mean absolute percentage error.	
RMSPE	Root-mean-square percentage error	
IAPWS-IF97	Industrial Formulation 1997 for water and steam properties	
ML	Machine learning	
UQ	Uncertainty quantification	
CFD	Computational fluid dynamics	
R_1 , R_2	Phenomenological response-pattern labels from the independent TRACE topology audit	response labels; not DNB/dryout
ARI	Adjusted Rand index used in response-topology stability auditing	dimensionless
BIC	Bayesian information criterion used in TRACE topology selection	dimensionless criterion
TRACE	Independent hierarchical multiscale response-field robustness audit used to test campaign effects and response topology	audit label

References

- [1] Todreas, N. E., & Kazimi, M. S. (2021). *Nuclear systems volume I: Thermal hydraulic fundamentals*. CRC Press.
- [2] Bruder, M., Bloch, G., Sattelmayer, T. (2016). Critical Heat Flux in Flow Boiling—Review of the Current Understanding and Experimental Approaches. *Heat Transfer Engineering*. <https://doi.org/10.1080/01457632.2016.1189274>
- [3] Lee, C., Mudawwar, I. (1988). A mechanistic critical heat flux model for subcooled flow boiling based on local bulk flow conditions. *International Journal of Multiphase Flow*. [https://doi.org/10.1016/0301-9322\(88\)90070-5](https://doi.org/10.1016/0301-9322(88)90070-5)
- [4] Weisman, J., Pei, B. (1983). Prediction of critical heat flux in flow boiling at low qualities. *International Journal of Heat and Mass Transfer*. [https://doi.org/10.1016/s0017-9310\(83\)80047-7](https://doi.org/10.1016/s0017-9310(83)80047-7)
- [5] Hejzlar, P., Todreas, N. (1996). Consideration of critical heat flux margin prediction by subcooled or low quality critical heat flux correlations. *Nuclear Engineering and Design*. [https://doi.org/10.1016/0029-5493\(95\)01169-2](https://doi.org/10.1016/0029-5493(95)01169-2)
- [6] Groeneveld, D., Shan, J., Vasić, A., Leung, L., Durmayaz, A., Yang, J., Cheng, S., Tanase, A. (2007). The 2006 CHF look-up table. *Nuclear Engineering and Design*. <https://doi.org/10.1016/j.nucengdes.2007.02.014>
- [7] Groeneveld, D. C. (2019). *Critical heat flux data used to generate the 2006 Groeneveld critical heat flux lookup tables*. United States Nuclear Regulatory Commission, Office of Nuclear Regulatory Research.
- [8] Bowring, R. W. (1972). A simple but accurate round tube, uniform heat flux, dryout correlation over the pressure range 0. 7-17MN/m² (100-2500psia). *AEW-R789*.
- [9] Biasi, L., Clerici, G., Garribba, S., Sala, R., & Tozzi, A. (1967). Studies on burnout. Part 3. A new correlation for round ducts and uniform heating and its comparison with world data. OSTI OAI (U.S. Department of Energy Office of Scientific and Technical Information).
- [10] Katto, Y., Ohno, H. (1984). An improved version of the generalized correlation of critical heat flux for the forced convective boiling in uniformly heated vertical tubes. *International Journal of Heat and Mass Transfer*. [https://doi.org/10.1016/0017-9310\(84\)90276-x](https://doi.org/10.1016/0017-9310(84)90276-x)
- [11] Mohammed Shah, M. (1987). Improved general correlation for critical heat flux during upflow in uniformly heated vertical tubes. *International Journal of Heat and Fluid Flow*. [https://doi.org/10.1016/0142-727x\(87\)90069-5](https://doi.org/10.1016/0142-727x(87)90069-5)
- [12] Tong, L. (1967). Prediction of departure from nucleate boiling for an axially non-uniform heat flux distribution. *Journal of Nuclear Energy*. [https://doi.org/10.1016/s0022-3107\(67\)90054-8](https://doi.org/10.1016/s0022-3107(67)90054-8)
- [13] Zhang, W., Hibiki, T., Mishima, K., Mi, Y. (2006). Correlation of critical heat flux for flow boiling of water in mini-channels. *International Journal of Heat and Mass Transfer*. <https://doi.org/10.1016/j.ijheatmasstransfer.2005.09.004>
- [14] Aksan, N., D'Auria, F. S., Groeneveld, D., Kirillov, P., Saha, D., Badulescu, A., & Cleveland, J. (2001). Thermohydraulic relationships for advanced water cooled reactors.
- [15] Celata, G., Cumo, M., Mariani, A. (1994). Assessment of correlations and models for the prediction of CHF in water subcooled flow boiling. *International Journal of Heat and Mass Transfer*. [https://doi.org/10.1016/0017-9310\(94\)90096-5](https://doi.org/10.1016/0017-9310(94)90096-5)
- [16] Hall, D., Mudawar, I. (2000). Critical heat flux (CHF) for water flow in tubes—II.. *International Journal of Heat and Mass Transfer*. [https://doi.org/10.1016/s0017-9310\(99\)00192-1](https://doi.org/10.1016/s0017-9310(99)00192-1)
- [17] Kim, S., Mudawar, I. (2013). Universal approach to predicting saturated flow boiling heat transfer in mini/micro-channels – Part I. Dryout incipience quality. *International Journal of Heat and Mass Transfer*. <https://doi.org/10.1016/j.ijheatmasstransfer.2013.04.016>
- [18] Liang, G., Mudawar, I. (2020). Review of channel flow boiling enhancement by surface modification, and instability suppression schemes. *International Journal of Heat and Mass Transfer*. <https://doi.org/10.1016/j.ijheatmasstransfer.2019.118864>
- [19] Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. <https://doi.org/10.5281/zenodo.19712698>
- [20] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection.

- [21] Roberts, D., Bahn, V., Ciuti, S., Boyce, M., Elith, J., Guillera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J., Schröder, B., Thuiller, W., Warton, D., Wintle, B., Hartig, F., Dormann, C. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*. <https://doi.org/10.1111/ecog.02881>
- [22] Varma, S., Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. <https://doi.org/10.1186/1471-2105-7-91>
- [23] Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. <https://doi.org/10.5555/1756006.1859921>
- [24] Zhao, X., Shirvan, K., Salko, R., Guo, F. (2020). On the prediction of critical heat flux using a physics-informed machine learning-aided framework. *Applied Thermal Engineering*. <https://doi.org/10.1016/j.applthermaleng.2019.114540>
- [25] Khalid, R., Ullah, A., Khan, A., Khan, A., Inayat, M. (2023). Comparison of Standalone and Hybrid Machine Learning Models for Prediction of Critical Heat Flux in Vertical Tubes. *Energies*. <https://doi.org/10.3390/en16073182>
- [26] Helmryd Grosfilley, E., Robertson, G., Soibam, J., Le Corre, J. (2024). Investigation of Machine Learning Regression Techniques to Predict Critical Heat Flux Over a Large Parameter Space. *Nuclear Technology*. <https://doi.org/10.1080/00295450.2024.2380580>
- [27] Furlong, A., Zhao, X., Salko, R., Wu, X. (2025). Physics-based hybrid machine learning for critical heat flux prediction with uncertainty quantification. *Applied Thermal Engineering*. <https://doi.org/10.1016/j.applthermaleng.2025.127447>
- [28] Mao, C., Jin, Y. (2024). Uncertainty quantification study of the physics-informed machine learning models for critical heat flux prediction. *Progress in Nuclear Energy*. <https://doi.org/10.1016/j.pnucene.2024.105097>
- [29] Alsafadi, F., Furlong, A., Wu, X. (2025). Predicting critical heat flux with uncertainty quantification and domain generalization using conditional variational autoencoders and deep neural networks. *Annals of Nuclear Energy*. <https://doi.org/10.1016/j.anucene.2025.111502>
- [30] Zhang, C., Wu, H., Tan, B., Zhao, J. (2026). Advancing nuclear energy safety via hybrid stacking models: Integrating physics-informed machine learning and traditional AI for critical heat flux prediction. *Progress in Nuclear Energy*. <https://doi.org/10.1016/j.pnucene.2026.106326>
- [31] Cazzola, M., Ghione, A., Sargentini, L., Nespoulous, J., & Finotello, R. (2026). Learning Complex Physical Regimes via Coverage-oriented Uncertainty Quantification: An application to the Critical Heat Flux. *arXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2602.21701>
- [32] Dunbar, C., Hwang, D., Duarte, J. (2026). Critical heat flux prediction for internally heated annuli with uniform power. *Nuclear Engineering and Design*. <https://doi.org/10.1016/j.nucengdes.2025.114685>
- [33] Karniadakis, G., Kevrekidis, I., Lu, L., Perdikaris, P., Wang, S., Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*. <https://doi.org/10.1038/s42254-021-00314-5>
- [34] Pang, Guofei, et al. "fPINNs: Fractional Physics-Informed Neural Networks." *SIAM Journal on Scientific Computing*, 2019. <https://doi.org/10.1137/18m1229845>.
- [35] Brunton, S., Noack, B., Koumoutsakos, P. (2020). Machine Learning for Fluid Mechanics. *Annual Review of Fluid Mechanics*. <https://doi.org/10.1146/annurev-fluid-010719-060214>
- [36] Duraisamy, K., Iaccarino, G., Xiao, H. (2019). Turbulence Modeling in the Age of Data. *Annual Review of Fluid Mechanics*. <https://doi.org/10.1146/annurev-fluid-010518-040547>
- [37] Delipei, G., Hou, J., Avramova, M., Rouxelin, P., Ivanov, K. (2021). Summary of comparative analysis and conclusions from OECD/NEA LWR-UAM benchmark Phase I. *Nuclear Engineering and Design*. <https://doi.org/10.1016/j.nucengdes.2021.111474>
- [38] Wu, X., Moloko, L., Bokov, P., Delipei, G., Kaizer, J., Ivanov, K. (2025). Uncertainty Quantification for Data-Driven Machine Learning Models in Nuclear Engineering Applications: Where We Are and What Do We Need?. *Nuclear Science and Engineering*. <https://doi.org/10.1080/00295639.2025.2552500>
- [39] Angelopoulos, A., Bates, S. (2023). Conformal Prediction: A Gentle Introduction. *Foundations and Trends® in Machine Learning*. <https://doi.org/10.1561/22000000101>
- [40] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.2017.1307116>
- [41] Barber, R., Candès, E., Ramdas, A., Tibshirani, R. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*. <https://doi.org/10.1214/20-aos1965>
- [42] Tibshirani, R. J., Foygel Barber, R., Candès, E., & Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.
- [43] Wilks, S. (1941). Determination of Sample Sizes for Setting Tolerance Limits. *The Annals of Mathematical Statistics*. <https://doi.org/10.1214/aoms/1177731788>
- [44] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NIPS*. <https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html>
- [45] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- [46] Aas, K., Jullum, M., Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*. <https://doi.org/10.1016/j.artint.2021.103502>
- [47] Ghorbani, A., & Zou, J. (2019, May). Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning* (pp. 2242-2251). PMLR.
- [48] Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., ... & Spanos, C. J. (2019, April). Towards efficient data valuation based on the Shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics* (pp. 1167-1176). PMLR.
- [49] Pan, S., Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/tkde.2009.191>
- [50] Wang, Z., Dai, Z., Póczos, B., & Carbonell, J. (2019, June). Characterizing and avoiding negative transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11285-11294). IEEE.
- [51] Sugiyama, M., Krauledat, M., & Müller, K. (2007). Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.*. <https://dl.acm.org/doi/10.5555/1314498.1390324>
- [52] Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885-1894). Pmlr.
- [53] Wright, L., Davidson, S. (2020). How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*. <https://doi.org/10.1186/s40323-020-00147-4>

- [54] Rasheed, A., San, O., Kvamsdal, T. (2020). Digital Twin: Values, Challenges and Enablers From a Modeling Perspective. IEEE Access. <https://doi.org/10.1109/access.2020.2970143>
- [55] Thelen, A., Zhang, X., Fink, O., Lu, Y., Ghosh, S., Youn, B., Todd, M., Mahadevan, S., Hu, C., Hu, Z. (2022). A comprehensive review of digital twin — part 1: modeling and twinning enabling technologies. Structural and Multidisciplinary Optimization. <https://doi.org/10.1007/s00158-022-03425-4>
- [56] Wagner, W., & Kretzschmar, H. J. (2008). *International steam tables: Properties of water and steam based on the industrial formulation IAPWS-IF97*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- [57] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: Springer.
- [58] Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. J. Mach. Learn. Res.. <https://dl.acm.org/doi/10.5555/2503308.2188395>
- [59] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res.. <https://dl.acm.org/doi/10.5555/1953048.2078195>
- [60] Geurts, P., Ernst, D., Wehenkel, L. (2006). Extremely randomized trees. Machine Learning. <https://doi.org/10.1007/s10994-006-6226-1>
- [61] Friedman, J. (2001). Greedy function approximation: A gradient boosting machine.. The Annals of Statistics. <https://doi.org/10.1214/aos/1013203451>
- [62] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. NIPS. <https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- [63] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- [64] Breiman, L. (2001). Random Forests. Machine Learning. <https://doi.org/10.1023/a:1010933404324>
- [65] Efron, B., Tibshirani, R. (1993). An Introduction to the Bootstrap. <https://doi.org/10.1007/978-1-4899-4541-9>
- [66] Roy, C., Oberkampf, W. (2011). A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. Computer Methods in Applied Mechanics and Engineering. <https://doi.org/10.1016/j.cma.2011.03.016>