When More Data Is Not Enough: Identifying and Repairing Support Gaps in Mechanism-Resolved Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Accurate marginal prediction does not necessarily determine the predictive law for a specified combination of factors. We quantify this ambiguity with the Support-Induced Composition Radius (SICR), the minimax information radius of target laws compatible with the population training experiment. Positive SICR gives an architecture-independent lower bound on worst-case target excess log loss at every sample size. Within a specified interaction family, we prove that SICR vanishes exactly when the target interaction feature lies in the training row span, under composition-wise context overlap and anchored reference and atomic effects. This characterization yields target-span acquisition (TSA), a greedy rule that selects compositions to reduce the target residual outside the training span. Controlled experiments show a persistent error plateau when sampling from the same support and its removal when the span is expanded. Across 200 six-factor design tasks, three TSA additions reduce the mean normalized squared target residual to 0.0023, compared with 0.5057 for regularized D-optimal design. In a retrospective UCI Bike Sharing audit, TSA reduces target RMSLE averaged across four predictors by 11.4% relative to the best non-targeted policy at 200 added records. An omitted-interaction stress test evaluates the certificate's sensitivity to the chosen interaction basis. Together, these results provide a target-specific workflow for diagnosing support gaps, choosing informative additions, and checking the structural assumptions behind extrapolation.

1 Introduction

Forecasting systems often estimate a marginal predictive law $p(Y \mid H)$ from a history or context H. Many scientific and operational queries instead specify a combination of factors, regimes, or control inputs z and request $p(Y \mid H, z)$. We call prediction at this resolution *mechanism-resolved forecasting*. The composition defines the predictive query; a causal interpretation requires additional assumptions.

A training experiment can identify its marginal law while leaving several target laws compatible with exactly the same observations. More samples can reduce estimation error, but cannot distinguish models that induce the same population training law. The practical question is therefore which new compositions would supply the missing information.

We answer this question in two steps. First, we define the target-law identified set of atom-preserving, training-equivalent models and quantify its ambiguity by the *Support-Induced Composition Radius* (SICR). Positive SICR gives a lower bound on worst-case target excess log loss for every sample size and every learner using only the training experiment. Second, within a specified interaction family, we show exactly when this ambiguity vanishes:

$$\rho_C^{\nu}(z^\star) = 0 \quad \Longleftrightarrow \quad \chi(z^\star) \in \mathrm{rowspan}(X_{\mathrm{train}}).$$

Here χ encodes residual interactions, X_{train} contains their training-support values, and ν specifies the target context distribution. The equivalence assumes anchored reference and atomic effects, composition-wise context overlap, and locally admissible interaction perturbations. An unobserved target law can therefore be identified when its interaction feature is already in the span. Repeated sampling from the existing support cannot supply a missing direction.

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

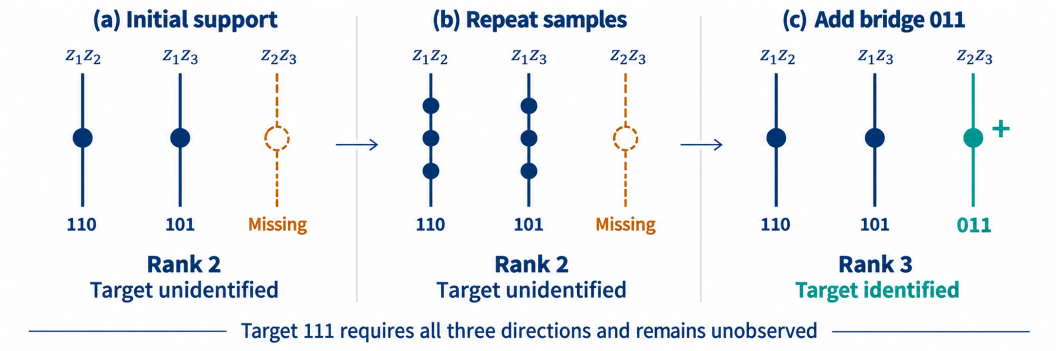


Figure 1: **Support repair in a pairwise interaction model.** With anchored reference and atomic effects, $\chi(z) = (z_1z_2, z_1z_3, z_2z_3)$. In each panel, the three channels show interaction directions; filled nodes represent schematic observations. Initial support $\{000, 100, 010, 001, 110, 101\}$ spans only the first two directions. Repeated sampling preserves interaction rank 2; bridge 011 supplies the third direction and identifies the law at unobserved target 111. Reference and atomic compositions have zero interaction features and are omitted.

This geometry gives *target-span acquisition* (TSA). The rule selects the candidate that removes the largest fraction of the squared target residual outside the current training span. Selection uses composition features without candidate labels (Figure 1).

Our experiments distinguish three questions: whether target-law ambiguity exists, whether changing support removes it in a controlled family, and whether target-aligned acquisition improves finite-sample prediction. We combine exhaustive support enumeration, stochastic dynamics, 200 multi-candidate design tasks, an omitted-interaction stress test, and a retrospective UCI Bike Sharing audit. The real-data study compares four predictors and four acquisition policies on the held-out composition "2012 $\times$ rush hour $\times$ adverse weather."

Our contributions are fourfold:

1. **Target-law ambiguity.** SICR connects a population identified set to an architecture-independent, all-N lower bound on worst-case target excess log loss.
2. **An exact support certificate.** Within the specified interaction family, target identification is equivalent to row-span coverage, including targets absent from training support.
3. **Targeted support repair.** TSA converts the certificate into a sequential acquisition rule; the omitted-interaction test provides a check on the maintained interaction basis.
4. **Controlled and real-data evidence.** Equal-budget interventions, multi-candidate design comparisons, and predictor-specific budget curves separate structural identification from finite-sample prediction gains.

2 Problem Setup

Let H denote the observed history or context available to a forecaster, $Z \in \mathcal{Z}$ a composition of active factors, and Y the outcome to be predicted. A model Q specifies a composition-conditioned predictive law $p_Q(y \mid h, z)$. The coordinates and interpretation of Z specify the query. Throughout, identification means that the training experiment uniquely determines the corresponding predictive law.

We distinguish two forecasting tasks. **Marginal forecasting (Task M)** predicts

$$H \longmapsto p_Q(Y \mid H),$$

where the composition is averaged according to the observational mixture represented by the experiment. For discrete compositions,

$$p_Q(y \mid h) = \sum_{z \in \mathcal{Z}} p_Q(y \mid h, z)\, p_Q(z \mid h),$$

with the analogous integral for continuous Z. **Mechanism-resolved forecasting (Task R)** instead predicts

$$(H, z) \longmapsto p_Q(Y \mid H, z)$$

for a specified composition z. The marginal law in Task M can therefore be identified even when the composition-conditioned laws in Task R are not.

The training experiment need not contain every composition at which a forecast may later be requested. We write $S_{\mathrm{train}} \subseteq \mathcal{Z}$ for the **population training composition support**: the set of compositions represented by the population experiment, not merely those observed in one finite sample. Our target is a composition $z^\star$, which may lie outside S_{train}. Target predictions are evaluated under a fixed context distribution ν. The general lower bound applies to this chosen measure. For the span characterization in Section 5, we additionally require that the training experiment identify each training-support conditional law for ν-almost every context; overlap with the pooled context support alone is insufficient.

The key question is whether the training experiment uniquely determines the law at $z^\star$, including when that composition is unobserved. To make this precise, let $\mathbb{P}^Q_{\mathrm{tr},N}$ denote the distribution of an N-sample training dataset generated by the fixed training experiment under model Q. We call two models **training-equivalent** when they induce the same observable training-data law:

$$\mathbb{P}^Q_{\mathrm{tr},N} = \mathbb{P}^{Q'}_{\mathrm{tr},N} \qquad \text{for every } N.$$

Equivalently, no amount of additional sampling from that unchanged experiment can statistically distinguish them.

Training equivalence alone can include changes of latent representation that also change the meaning of the composition coordinates. Such comparisons would not define a meaningful mechanism-resolved target. We therefore compare **atom-preserving training-equivalent models**: models that preserve the atomic effects and the meaning of the composition coordinates, up to the allowed gauge symmetries. We denote this restricted equivalence class around the data-generating model $P^\star$ by

$$\mathfrak{E}_{\mathrm{atom}}(P^\star).$$

The subscript C used below denotes this fixed composition query and maintained model class. Section 5 and Appendix A.4 specify the reference law and normalization convention for the structured family.

The distinction between marginal prediction and target-law identification is thus a distinction between two statistical queries. We next quantify how much the population experiment leaves unresolved about the latter.

3 Target-Law Non-Identification

Fix a reference data-generating model $P^\star$. The atom-preserving equivalence class $\mathfrak{E}_{\mathrm{atom}}(P^\star)$ contains models with the same observable population training law as $P^\star$ and the same atomic effects and composition semantics needed to interpret the target query. Their compatible predictions at $z^\star$ form the target-law identified set

$$\mathcal{T}_C(z^\star) = \{p_Q(Y \mid H, z^\star) : Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)\}.$$

For a candidate target predictor $q(\cdot \mid H)$, define

$$R^\nu_Q(q; z^\star) = \mathbb{E}_{H\sim\nu}[D_{\mathrm{KL}}(p_Q(\cdot \mid H, z^\star) \| q(\cdot \mid H))],$$

and

$$\rho^\nu_C(z^\star) = \inf_q \sup_{Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)} R^\nu_Q(q; z^\star).$$

We call the minimax information radius of this target-law set the Support-Induced Composition Radius (SICR). SICR measures target-law disagreement that remains even when the population training experiment is known exactly, thereby separating structural ambiguity from finite-sample estimation uncertainty.

Under the conditions formalized in Appendix A, $\rho^\nu_C(z^\star) = 0$ exactly when all models in $\mathfrak{E}_{\mathrm{atom}}(P^\star)$ induce the same target law for ν-almost every context. Two training-equivalent worlds that disagree

at $z^\star$ already provide a positive witness; Appendix A.2 gives the corresponding Jensen–Shannon lower bound.

We next translate this population ambiguity into a bound for finite-sample learners.

4 Sample-Size-Independent Impossibility

Consider any training-only learning algorithm $\mathcal{A}_N$ that receives D_N from the fixed training experiment and returns

$$\widehat{q}_N(\cdot \mid H, z^\star) = \mathcal{A}_N(D_N).$$

For $Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)$, let

$$\mathcal{L}_Q^\nu(\mathcal{A}_N; z^\star) = \mathbb{E}_{D_N}\mathbb{E}_{H\sim\nu}[D_{\mathrm{KL}}(p_Q(\cdot \mid H, z^\star)\|\widehat{q}_N(\cdot \mid H, z^\star))]\,,$$

where the training-data distribution is common to all training-equivalent worlds; for randomized learners, the expectation also includes world-independent internal randomness.

Theorem 1 (sample-size-independent target ambiguity). For every sample size N,

$$\inf_{\mathcal{A}_N} \sup_{Q\in\mathfrak{E}_{\mathrm{atom}}(P^\star)} \mathcal{L}_Q^\nu(\mathcal{A}_N; z^\star) \geq \rho_C^\nu(z^\star).$$

If $\rho_C^\nu(z^\star) > 0$, worst-case target excess log loss is therefore bounded away from zero for every N. Additional observations from the unchanged training experiment cannot remove this ambiguity.

The proof uses only the fact that training-equivalent worlds induce the same law for D_N. Dataset-dependent predictions can therefore be averaged under a distribution common to every candidate world, and convexity of KL in its second argument reduces the learner to a single target predictor whose worst-case loss is at least the information radius. Appendix A.3 gives the formal argument.

The theorem distinguishes estimation error from structural ambiguity. More samples can improve estimation of identified quantities, but cannot distinguish training-equivalent models with different target laws. This is compatible with accurate marginal forecasting because the lower bound concerns $p(Y \mid H, z^\star)$, not necessarily $p(Y \mid H)$.

Zero SICR means that the target law is identified. Estimating it from finite data is a separate question. The next section gives a concrete condition for identification.

5 Interaction-Span Characterization

We now specialize to a structured interaction family on a finite composition set. Take the anchored reference composition to be $z_0 = 0$, and let

$$r_z(y, h) = \mathcal{C}_h\left[\log\frac{p(y \mid h, z)}{p(y \mid h, z_0)}\right], \qquad r_z = \sum_{k=1}^{K} z_k r_k + \chi(z)^\top\beta,$$

where $\mathcal{C}_h$ removes outcome-independent additive normalization, r_k are atomic effects, $\chi(z) \in \mathbb{R}^d$ is a known residual interaction feature map satisfying $\chi(0) = 0$ and $\chi(e_k) = 0$ for each anchored atomic composition e_k, and β collects interaction effects. Appendix A.4 gives the quotient-space construction that makes this representation distributionally well defined.

The characterization fixes the factorization, composition coordinates, reference law $p(\cdot \mid h, 0)$, atomic effects, and correctly specified interaction map. It also requires *composition-wise context overlap*: training equivalence determines $p(\cdot \mid h, z)$ for every $z \in S_{\mathrm{train}}$ and ν-almost every h. For observed discrete compositions, a sufficient condition is $\nu \ll P_{\mathrm{tr}}(H \mid Z = z)$ for each such z, with a common design law. Finally, a nonzero bounded distributional effect must be available on a positive-ν-measure set, and sufficiently small perturbations along each interaction direction must remain admissible. Appendix A.4 states these conditions precisely. Let X_{train} have rows $\chi(z)^\top$ for $z \in S_{\mathrm{train}}$. Training-invisible interaction perturbations then satisfy $X_{\mathrm{train}}\Delta = 0$.

Theorem 2 (target-relevant interaction-span characterization). Under these assumptions,

$$\rho_C^\nu(z^\star) = 0 \quad\Longleftrightarrow\quad \chi(z^\star) \in \mathrm{rowspan}(X_{\mathrm{train}}).$$

The characterization follows from null-space duality applied to target-law variation. If $\chi(z^\star)$ lies in the row span, every interaction perturbation invisible on the training support is also invisible at the target. If it lies outside the row span, a null-space direction exists that leaves every training-support law unchanged while changing the target interaction contribution. Appendix A.4 proves both directions and handles gauge and admissibility details.

The target law can thus be identified even when the full interaction parameter β remains undetermined. For example, in the pairwise three-factor model, observing 110, 101, and 011 supplies all three interaction directions and identifies 111 without observing it.

The row-space condition is classical for estimability in linear models (Pukelsheim, 2006). Here the interaction coefficients are distributional effects modulo normalization, so the same algebra determines equality of target predictive laws and, through SICR, an all-N excess-risk obstruction.

This criterion separates sample count from support geometry. Repeated observations at already represented compositions can reduce estimation variance but do not change $\mathrm{rowspan}(X_{\mathrm{train}})$. A new composition can instead expand the span and remove a missing target-relevant direction.

5.1 From Certificate to Target-Span Acquisition

The characterization gives an acquisition rule that uses the declared interaction map but does not require fitting a predictor. Let $P_X = X_{\mathrm{train}}^{+} X_{\mathrm{train}}$ project onto the current row span and set

$$g_\star = (I - P_X)\chi(z^\star), \qquad g_c = (I - P_X)\chi(c)$$

for target $z^\star$ and candidate composition c. If $g_\star = 0$, the target is already covered. Otherwise, the fraction of its squared residual norm removed by adding c is

$$a(c; z^\star) = \begin{cases} \dfrac{\langle g_\star, g_c\rangle^2}{\|g_\star\|_2^2 \|g_c\|_2^2}, & \|g_c\|_2 > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Greedily maximizing $a(c; z^\star)$ is **target-span acquisition (TSA)**. It selects support before observing candidate labels; $a = 1$ removes the entire current target residual. After each selection, the projection is updated and another candidate can remove a different component of the remaining target residual. Rank-revealing QR or SVD costs $O(|S_{\mathrm{train}}|d^2 + d^3 + |\mathcal{C}|d^2)$ for candidate set $\mathcal{C}$. The score follows from the exact update $\|g_\star^{\mathrm{new}}\|_2^2 = \|g_\star\|_2^2 - \langle g_\star, g_c\rangle^2 / \|g_c\|_2^2$ for $g_c \neq 0$. TSA seeks target-span coverage. D-optimal design seeks global information, while classical c-optimal design minimizes the variance of an estimable linear functional. The ranking uses the Euclidean metric of the declared feature map; our binary experiments use unscaled monomials. Support additions change the experiment and shrink its equivalence class.

Section 6 tests the resulting predictions in controlled experiments. Section 7 evaluates TSA on real data.

6 Controlled Empirical Consequences

We first measure how often support leaves a target unidentified. We then test the predicted error plateau, the separation between Tasks M and R, and the effect of support expansion under equal budgets. Multi-candidate design and misspecification tests examine acquisition behavior and sensitivity to the interaction basis. Appendices B–E give the protocols and additional results.

6.1 Breadth of Support-Induced Ambiguity

Under the pairwise basis, exhaustive enumeration of all 525,344 target/support pairs for binary $K \in \{2, 3, 4\}$ finds 219,560 structurally ambiguous cases. The archived coordinate audit reports zero failures across 12,589,120 permutation checks (Appendix B).

6.2 More Samples on the Same Support Leave a Structural Plateau

For target 111, $S_{\mathrm{amb}} = \{000, 100, 010, 001, 110, 101\}$ misses one interaction direction and $S_{\mathrm{id}} = S_{\mathrm{amb}} \cup \{011\}$ restores it. In the stochastic system of Appendix C.1, the minimum-norm forcing

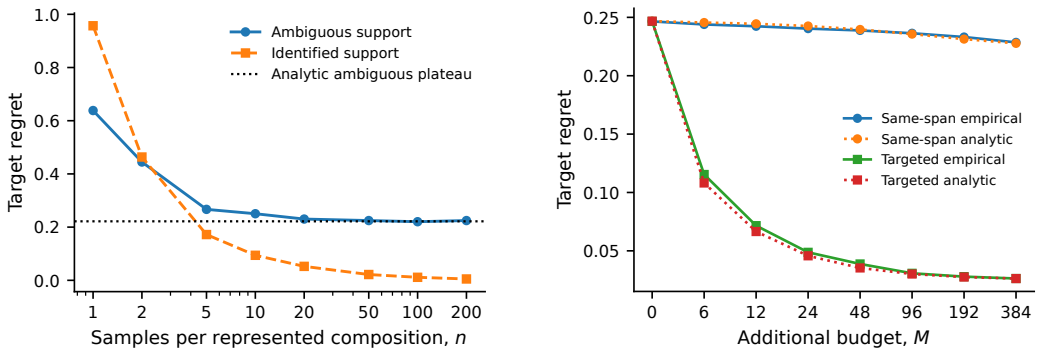


Figure 2: **Controlled consequences.** (a) In the plotted world $s = +1$, increasing samples on ambiguous support approaches the forcing estimator's analytic plateau, while restoring 011 removes it (200 seeds per point). (b) Under equal total budgets, same-span replication leaves the target direction unresolved; targeted 011 acquisition expands the span and follows the analytic curve (200 seeds per world and point).

estimator has an analytic ambiguous-support plateau of 0.2218; in the plotted world $s = +1$ at $n = 200$, seed-mean regret is 0.2247 under S_{amb} versus 0.0051 under S_{id} (Figure 2a). The plateau is the error of this estimator, not a numerical evaluation of SICR over unrestricted predictive laws.

6.3 Accurate Marginal Prediction with an Unidentified Target Law

To test whether marginal prediction can improve despite target-law ambiguity, we hide the composition labels in the construction of Appendix C.2. The six training compositions are S_{amb}, latent Z is uniform on this support, and the observable training law is the resulting mixture $P_M^\star$. Task M is evaluated by

$$R_M(\mu) = D_{\mathrm{KL}}(P_M^\star \| q_\mu).$$

Two training-equivalent worlds induce the same observable mixture but different target-111 laws. Marginal regret nevertheless falls from 2.1675 to 0.0095 at $n = 200$ unlabeled trajectories (Appendix C.2), demonstrating accurate marginal prediction despite an unidentified target law.

6.4 Equal-Budget Support Intervention

We next isolate the effect of support while holding the total data budget fixed. Both arms begin with 20 trajectories on each of the six ambiguous-support compositions, so

$$N_{\mathrm{base}} = 120, \qquad M \in \{0, 6, 12, 24, 48, 96, 192, 384\}, \qquad N_{\mathrm{total}} = 120 + M.$$

The same-span arm samples the additional M observations from the existing support S_{amb}, leaving interaction rank 2. The targeted arm allocates them to 011, which raises the rank to 3 for $M > 0$ and supplies the missing target-relevant direction for $z^\star = 111$.

At $M = 384$, same-span and targeted mean regrets are 0.2286 and 0.0263; their difference of 0.2023 is close to the analytic prediction of 0.2015. Across all 32 cells, the standardized discrepancy is below the reference threshold ($T_D = 0.966 < 3.163$; Appendix C.3 and Figure 2b).

The unique bridge above makes a rank-gain rule coincide with TSA. In 200 six-factor tasks with many candidates, TSA instead reduces the mean normalized squared target residual after three additions to 0.0023, versus 0.5057 for D-optimal design, and reaches zero in all tasks by budget four. Appendix C.4 reports all acquisition curves and identification rates.

6.5 Interaction Misspecification Stress Test

Theorem 2 is conditional on the correctness of χ. We therefore construct a falsification test in which the maintained basis is pairwise but the true Gaussian mean contains an omitted term $\gamma z_1 z_2 z_3$ with

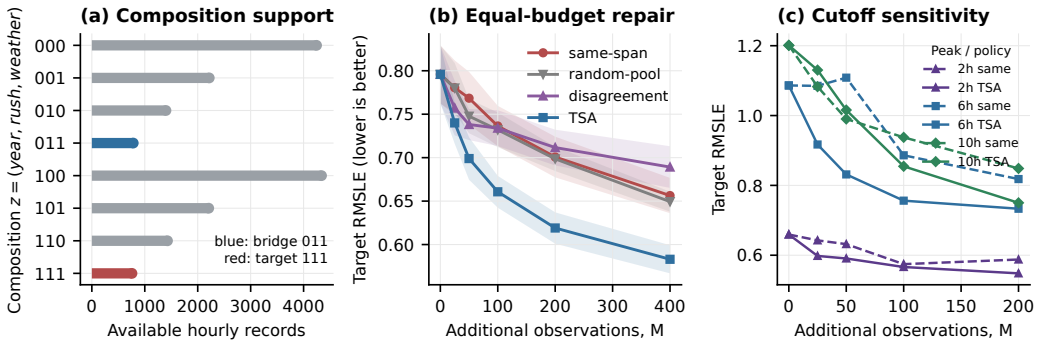


Figure 3: **Real-world support audit on the UCI Bike Sharing source table.** (a) Composition counts. (b) Mean target RMSLE across four forecasters and 10 seeds for same-span, random-pool, model-disagreement, and TSA acquisition; bands are 95% intervals over seed-wise model averages. (c) Same-span versus TSA HistGradientBoosting curves under three rush-hour definitions (five seeds).

unit observation variance. Every non-target binary composition has zero three-way feature, while $z^\star = 111$ has value one. After adding 011, the pairwise audit reports zero target residual, yet the omitted term creates target excess log loss $\gamma^2/2$. An audit using the enriched cubic basis detects the remaining support gap: the target lies outside the observed span (Figure 5, Appendix D). Additional observations on these non-target compositions leave γ unconstrained; resolving it requires new information or an additional structural assumption.

7 Real-World Audit: Bike-Sharing Demand

We use 17,379 UCI Bike Sharing hourly records (Fanaee-T, 2013; Fanaee-T & Gama, 2014); the same source dataset was used to construct Kaggle's Bike Sharing Demand benchmark (Cukierski, 2014). The identity $\texttt{cnt} = \texttt{casual} + \texttt{registered}$ holds on every UCI row, so using these post-outcome components gives RMSLE 0. We exclude both post-outcome fields from every reported predictor. Define $z = (\text{year 2012}, \text{rush hour}, \text{adverse weather})$, with rush hours $\{7, 8, 9, 16, 17, 18\}$ and adverse weather defined by $\texttt{weathersit} \geq 2$ (including mist/cloud). We hold out target 111 (758 records) and initialize with 100 rows from each of the six compositions excluding 011 and 111.

The four equal-budget acquisition policies are *same-span* (unused rows from the existing support), *random-pool* (uniformly sampled unused non-target rows), *model-disagreement* (rows with the largest cross-model predictive disagreement), and TSA (2011 rush-hour/adverse-weather bridge rows, 011). Selectors use no candidate or target outcomes; the disagreement selector is fitted using initial-support labels. TSA places the full budget on the bridge, raising pairwise rank from 2 to 3. Random-pool and disagreement acquisition can also reach rank 3 whenever they select at least one bridge row; the policies differ in how much data they allocate there.

This retrospective audit evaluates TSA with flexible predictors. Their full covariates include variables defining Z, so the pairwise calculation is not itself a certificate that their conditional target laws satisfy Theorem 2. Appendix E gives provenance and protocol.

7.1 Baselines, Ablations, and Efficiency

Conventional 80/20 random splits give RMSLE 0.306–0.594; the unadapted support holdout gives 0.556–0.890. These results use different test distributions and training sizes (Table 1). At $M = 200$, TSA obtains 0.619 mean RMSLE versus 0.699 for random-pool, an 11.4% reduction; the paired seed-wise difference is 0.0795 with descriptive 95% interval $[0.0566, 0.1025]$. Versus same-span, per-model gains are 15.2% (Ridge), 19.6% (random forest), 5.6% (HistGradientBoosting), and 1.3% (MLP). At $M = 400$, TSA reduces mean RMSLE by 10.2% relative to the strongest non-targeted policy (0.583 versus 0.649). Model-specific gains vary with budget; Appendix Figure 7 gives the full curves, including the MLP reversal at $M = 400$. Across three rush-hour definitions, TSA

outperforms same-span at $M = 200$. In the archived timing experiment, the $d = 3$ audit takes $57.8\,\mu$s, compared with 0.012–0.808s for model fitting (Table 1).

8 Related Work

Partial identification and minimax ambiguity. Our target-law identified set follows the partial-identification view that data and maintained assumptions may determine a set of admissible population quantities rather than a point (Manski, 2003; Tamer, 2010). We use the standard minimax information-radius construction to quantify this residual target-law ambiguity. Related work on partially identified distributional robustness studies population minimax risk under heterogeneous environments (Kostin et al., 2024). Our object differs: the training experiment is fixed, and we ask whether training-equivalent models disagree on one specified mechanism-resolved target law. SICR quantifies this target-law ambiguity, and Theorem 1 translates it into a lower bound on excess log loss at every sample size.

Latent identifiability and compositional generalization. Prior work studies conditions under which latent components or representations can be identified and recombined beyond observed combinations. This includes object-centric compositional generalization (Wiedemer et al., 2024), interaction-based disentanglement conditions (Brady et al., 2025), richer hierarchical or causal compositional structure (Kong et al., 2026), and distinctions between statistical and structural representation identifiability (Nelson et al., 2026). Our characterization is conditional on anchored composition coordinates, fixed or identified atomic semantics, and a specified interaction family. It therefore concerns identification of a particular target predictive law, not generic latent-representation recovery.

Perturbation extrapolation and support coverage. Work on perturbation extrapolation similarly emphasizes that the support of observed perturbations matters for generalization; for example, von Kügelgen et al. (2026) derive guarantees for unseen perturbations expressible as linear combinations of observed ones under an additive latent-shift model. Kernel analyses show that even explicitly compositional representations retain training-statistics-dependent failure modes (Lippl & Stachenfeld, 2025), and functional-equivalence analyses characterize pattern matching and its limits in compositional tasks (Chang et al., 2026). We instead begin with a fixed mechanism-resolved query and ask whether the population training support identifies its law. This yields a sample-size-independent non-identification result, a target-specific certificate, and an equal-budget comparison between additions that preserve or expand the relevant span.

Estimability and design. Classical linear-model theory relates estimability to row-space membership. D-optimal design seeks global information; c-optimal design instead minimizes the variance of a specified estimable linear functional (Pukelsheim, 2006). TSA addresses the preceding support question: which candidate most reduces the target residual outside the training span? Its objective is geometric coverage rather than finite-sample variance. Our contribution links this support question to equality of normalized target laws and the SICR risk bound.

Expressive predictors and external information. Transformers and diffusion Transformers provide expressive conditional model classes (Vaswani et al., 2017; Peebles & Xie, 2023). Theorem 1 applies to any such class when the learner uses only the fixed training experiment. Pretraining can supply additional information and thereby change the relevant equivalence class. Applying a span audit in a learned representation further requires a validated correspondence between its coordinates and the declared composition query.

9 Limitations

Structural assumptions. The exact certificate is relative to the specified interaction family, anchored reference and atomic effects, composition-wise context overlap, and locally admissible perturbations. The omitted-interaction test shows why checking the basis matters. Noisy or unavailable composition labels, feedback, and context shift require additional modeling.

Acquisition and finite samples. Exact population rank abstracts from rarity, numerical conditioning, acquisition cost, and noise. TSA is greedy and target-specific; its scores depend on feature scaling. Once the target span is covered, variance-aware allocation remains a separate design problem.

Empirical scope. The multi-candidate benchmark measures acquisition geometry without response noise, and the real-data evidence comes from one retrospective composition holdout. Further evaluation should cover additional domains and prospective acquisition.

10 Conclusion

Accurate marginal prediction need not identify the predictive law at a specified composition. SICR quantifies this ambiguity and bounds worst-case target excess log loss at every sample size. Within the stated interaction family, the ambiguity vanishes exactly when the target feature lies in the training row span. TSA uses this criterion to select compositions that reduce the uncovered target residual. The experiments show how support expansion can remove structural error and clarify how the interaction basis and finite-sample estimation affect predictive performance. The resulting workflow connects a target-specific support audit to acquisition and predictive evaluation.

Reproducibility Statement

The manuscript source, code, and experimental records are available anonymously at `https://anonymous.4open.science/r/support-gap-forecasting-D077/`. Appendix A provides the proofs, Appendix B specifies the support enumeration, and Appendices C–E detail the experimental protocols, data provenance, model settings, acquisition budgets, and evaluation procedures. The experimental pipeline produces 840 model/acquisition records, 135 cutoff-sensitivity records, and 5,600 multi-candidate design records. All numerical experiments are CPU-only.

AI Use Statement

AI tools were used to polish the language of this manuscript. The authors reviewed the final text and take full responsibility for its content.

Ethics Statement

The empirical study uses a public, aggregate bike-demand dataset released under CC BY 4.0; it introduces no new human-subject data collection and uses no personal identifiers. Appendix E.1 provides the data source, license, and file hash.

REFERENCES

Jack Brady, Julius von Kügelgen, Sébastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cC1l0TU836.

Hoyeon Chang, Jinho Park, Hanseul Cho, Sohee Yang, Miyoung Ko, Hyeonbin Hwang, Seungpil Won, Dohaeng Lee, Youbin Ahn, and Minjoon Seo. Characterizing pattern matching and its limits on compositional task structures. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/29f1595d10229ee89dcc8abbbd9c5d87-Abstract-Conference.html.

Will Cukierski. Bike sharing demand. Kaggle competition, https://www.kaggle.com/competitions/bike-sharing-demand, 2014. Accessed 2026-09-24.

Hadi Fanaee-T. Bike sharing. UCI Machine Learning Repository, https://doi.org/10.24432/C5W894, 2013. CC BY 4.0.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014. doi: 10.1007/s13748-013-0040-3.

Lingjing Kong, Shaoan Xie, Yang Jiao, Yetian Chen, Yanhui Guo, Simone Shao, Yan Gao, Guangyi Chen, and Kun Zhang. Learning by analogy: A causal framework for composition generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Julia Kostin, Nicola Gnecco, and Fanny Yang. Achievable distributional robustness when the robust risk is only partially identified. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-2667.

Samuel Lippl and Kimberly Stachenfeld. When does compositional structure yield compositional generalization? a kernel theory. In *International Conference on Learning Representations*, 2025.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer Series in Statistics. Springer, 2003. doi: 10.1007/b97478.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *International Conference on Learning Representations*, 2026.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, 2006. doi: 10.1137/1.9780898719109.

Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Julius von Kügelgen, Jakob Ketterer, Michael Vollenweider, Michael Scholkemper, Xinwei Shen, Nicolai Meinshausen, and Jonas Peters. Extrapolation guarantees for perturbation modeling under the additive latent shift assumption. *arXiv preprint arXiv:2504.18522v3*, 2026. URL https://arxiv.org/abs/2504.18522v3.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *International Conference on Learning Representations*, 2024.

A Formal Results and Proofs

A.1 The Log-Loss Value of Composition Information

The predictive value of observing a composition has a standard information-theoretic expression.

Let $P^\star$ denote the joint distribution of (H, Z, Y). Under logarithmic loss, the Bayes-optimal predictor that uses only H is $p^\star(y \mid h)$, whereas the Bayes-optimal predictor that additionally observes Z is $p^\star(y \mid h, z)$. When these entropies are finite, their optimal expected log losses are respectively $\mathcal{R}^\star_M = H(Y \mid H)$ and $\mathcal{R}^\star_R = H(Y \mid H, Z)$.

Proposition A.1 (log-loss value of composition information). Whenever the relevant conditional entropies are finite, $\mathcal{R}^\star_M - \mathcal{R}^\star_R = I(Y; Z \mid H)$.

Proof. By the definition of conditional mutual information, $I(Y; Z \mid H) = H(Y \mid H) - H(Y \mid H, Z)$. Substituting the Bayes risks above gives the stated identity. □

This identity describes the value of Z when the conditional laws are available. Whether a training experiment identifies those laws is the separate question addressed next.

A.2 SICR and Target-Law Identification

We now make precise the identification interpretation used in Section 3. Fix a target composition $z^\star$, a context measure ν, and the atom-preserving training-equivalence class $\mathfrak{E}_{\mathrm{atom}}(P^\star)$. For each $Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)$, write $P^h_Q = p_Q(\cdot \mid h, z^\star)$ for the target predictive law at context h. Recall that

$$\rho^\nu_C(z^\star) = \inf_q \sup_{Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)} \mathbb{E}_{H\sim\nu}\left[D_{\mathrm{KL}}\left(P^H_Q \| q(\cdot \mid H)\right)\right].$$

We assume throughout this subsection that the conditional probability kernels are measurable and that the displayed divergences are defined in the extended nonnegative reals. These regularity conditions are sufficient for the zero-radius characterization below.

Proposition A.2 (SICR characterizes target-law identification). $\rho^\nu_C(z^\star) = 0$ if and only if each model $Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)$ agrees with the reference model on the target predictive law ν-almost everywhere (the null set may depend on Q).

Proof. Suppose first that every model induces the same target law P^h_0 for ν-almost every h. Choosing $q(\cdot \mid h) = P^h_0$ gives zero KL divergence for every admissible model, hence $\rho^\nu_C(z^\star) = 0$.

For the converse, suppose two models Q_0, Q_1 disagree on their target laws on a set of contexts having positive ν-measure. Restricting the supremum to these two worlds and using $\max\{a, b\} \geq (a + b)/2$ gives

$$\rho^\nu_C(z^\star) \geq \inf_q \frac{1}{2}\left[R^\nu_{Q_0}(q; z^\star) + R^\nu_{Q_1}(q; z^\star)\right].$$

For each context h, define $M^h = (P^h_{Q_0} + P^h_{Q_1})/2$. The KL barycenter identity gives

$$\inf_q \left[\frac{1}{2} D_{\mathrm{KL}}(P^h_{Q_0} \| q) + \frac{1}{2} D_{\mathrm{KL}}(P^h_{Q_1} \| q)\right] = \mathrm{JS}(P^h_{Q_0}, P^h_{Q_1}),$$

where JS is the equal-weight Jensen–Shannon divergence. Therefore

$$\rho^\nu_C(z^\star) \geq \mathbb{E}_{H\sim\nu}[\mathrm{JS}(P^H_{Q_0}, P^H_{Q_1})].$$

Jensen–Shannon divergence is nonnegative and equals zero exactly when its arguments agree. Since the worlds disagree on a positive-measure set, the expectation is strictly positive, proving the converse. □

The proof also yields the two-world lower bound used in the main text:

$$\rho^\nu_C(z^\star) \geq \mathbb{E}_{H\sim\nu}\left[\mathrm{JS}\left(p_{Q_0}(\cdot \mid H, z^\star), p_{Q_1}(\cdot \mid H, z^\star)\right)\right].$$

Thus one explicit pair of atom-preserving training-equivalent worlds is already sufficient to certify positive structural ambiguity whenever their target laws differ on a set of positive ν-measure.

The argument concerns equality of predictive laws and is independent of the structured interaction assumptions used in Theorem 2.

A.3 Proof of the Sample-Size-Independent Impossibility

We now prove Theorem 1 from Section 4. The argument shows why predictions fitted to a dataset cannot resolve ambiguity left by the population training experiment.

Fix a sample size N and any training-only learning rule $\mathcal{A}_N$. For a randomized rule, include its independent random seed in D_N; the augmented input law is still common to all worlds. For a realized input D_N, write $\widehat{q}_{N,D_N}(\cdot \mid h, z^\star) = \mathcal{A}_N(D_N)(\cdot \mid h, z^\star)$. Every $Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)$ is training-equivalent to $P^\star$, so the distribution of D_N is the same under every candidate world and equals $\mathbb{P}^{P^\star}_{\mathrm{tr},N}$.

Define the dataset-averaged predictive kernel $\bar{q}_N(\cdot \mid h, z^\star) = \mathbb{E}_{D_N \sim \mathbb{P}^{P^\star}_{\mathrm{tr},N}}[\widehat{q}_{N,D_N}(\cdot \mid h, z^\star)]$. Under the standing measurability assumptions this pointwise mixture is an admissible predictive probability kernel.

Fix $Q \in \mathfrak{E}_{\mathrm{atom}}(P^\star)$ and context h, and abbreviate $P^h_Q = p_Q(\cdot \mid h, z^\star)$. The map $q \mapsto D_{\mathrm{KL}}(P^h_Q \| q)$ is convex in its second argument. Jensen's inequality therefore gives

$$D_{\mathrm{KL}}(P^h_Q \| \bar{q}_N) \le \mathbb{E}_{D_N}[D_{\mathrm{KL}}(P^h_Q \| \widehat{q}_{N,D_N})].$$

Averaging over $H \sim \nu$ yields $R^\nu_Q(\bar{q}_N; z^\star) \le \mathcal{L}^\nu_Q(\mathcal{A}_N; z^\star)$. Taking the supremum over Q and using the SICR definition gives

$$\rho^\nu_C(z^\star) \le \sup_Q R^\nu_Q(\bar{q}_N; z^\star) \le \sup_Q \mathcal{L}^\nu_Q(\mathcal{A}_N; z^\star).$$

Because $\mathcal{A}_N$ was arbitrary, $\inf_{\mathcal{A}_N} \sup_Q \mathcal{L}^\nu_Q(\mathcal{A}_N; z^\star) \ge \rho^\nu_C(z^\star)$. This is Theorem 1. □

The essential step is the common law of the learner's input: averaging its output cannot introduce information distinguishing the candidate worlds.

A.4 Interaction-Span Characterization and Gauge Details

We now prove Theorem 2 and make explicit the quotient structure implicit in the main-text notation. Interaction coefficients take values in a space of distributional effects; only the target predictive law, not a particular parameter representation, is the object of the theorem.

A.4.1 Distributional effects modulo normalization

Fix context h and reference composition z_0. For composition z, let $\ell_z(y, h) = \log[p(y \mid h, z)/p(y \mid h, z_0)]$. Assume the modeled laws have strictly positive densities on a common outcome support and work in the quotient $\mathcal{V}_h = \mathcal{F}_h/\mathcal{N}_h$, where $\mathcal{N}_h = \{c(h)\mathbf{1} : c(h) \in \mathbb{R}\}$ contains outcome-independent additive functions. The centering operator $\mathcal{C}_h$ chooses a representative. With a fixed reference law, equality in this quotient implies equality of normalized predictive laws: the two log ratios differ only by a constant, and normalization forces that constant to be zero.

A.4.2 Structured interaction family

Under the assumptions of Theorem 2, the anchored reference is $z_0 = 0$, and χ is a residual interaction map satisfying $\chi(0) = 0$ and $\chi(e_k) = 0$ for each anchored atomic composition e_k. Hence perturbations of β preserve the reference and atomic effects. The structured family is $r_z = \sum_{k=1}^K z_k r_k + \chi(z)^\top \beta$, where $r_k \in \mathcal{V}$, $\chi(z) \in \mathbb{R}^d$, and $\beta \in \mathcal{V}^d$. The matrix X_{train} has rows $\chi(z)^\top$ for $z \in S_{\mathrm{train}}$ and acts only on the finite-dimensional interaction index. For two models, $\Delta = \beta_{Q'} - \beta_Q \in \mathcal{V}^d$ and their effect difference is $r^{Q'}_z - r^Q_z = \chi(z)^\top \Delta$. Training equivalence implies $X_{\mathrm{train}}\Delta = 0$ componentwise in $\mathcal{V}$ at the evaluation contexts.

Here $\mathcal{V}$ denotes measurable context-indexed quotient effects, with equality modulo ν-null sets. The reference law and atomic effects are shared across the equivalence class. The observable training law is determined by a fixed design and the composition-conditioned kernels on its support, as in independent draws with observed (H, Z). Thus preserving these kernels at every training context preserves the entire training experiment. We restrict this characterization to a finite composition set and an experiment whose observable law determines the composition-conditioned laws on S_{train}

at the evaluation contexts. For observed discrete Z, this follows from $\nu \ll P_{\mathrm{tr}}(H \mid Z = z)$ for every training composition and a fixed sampling law for (H, Z). In a latent-mixture experiment this property requires a separate identification argument; mixture equality alone does not imply $X_{\mathrm{train}}\Delta = 0$.

For the converse, assume there exists a bounded measurable effect $g(y, h)$ that is nonconstant in y under the reference law on a set of contexts of positive ν-measure. For every $a \in \mathbb{R}^d$, the family admits the perturbation $\beta_j \mapsto \beta_j + \epsilon a_j g$ for sufficiently small nonzero ϵ. It is realized by the normalized exponential tilt

$$p_\epsilon(y \mid h, z) = \frac{p(y \mid h, z) \exp\{\epsilon \chi(z)^\top a\, g(y, h)\}}{\int p(u \mid h, z) \exp\{\epsilon \chi(z)^\top a\, g(u, h)\}\, du}.$$

Boundedness makes the normalizer finite. The vanishing interaction features at 0 and e_k preserve the reference and atomic laws; a null-space a also preserves every training-support law. This states explicitly the local richness needed for the necessity direction.

A.4.3 Proof of Theorem 2

We prove $\rho^\nu_C(z^\star) = 0 \iff \chi(z^\star) \in \mathrm{rowspan}(X_{\mathrm{train}})$. If the target feature lies in the row span, write $\chi(z^\star) = X_{\mathrm{train}}^\top \lambda$. For every training-invisible Δ, $\chi(z^\star)^\top \Delta = \lambda^\top X_{\mathrm{train}} \Delta = 0$. Quotient injectivity then gives equality of target predictive laws for every admissible pair, so Proposition A.2 yields zero SICR.

Conversely, if $\chi(z^\star)$ is outside the row span, finite-dimensional orthogonality gives $a \in \mathrm{Null}(X_{\mathrm{train}})$ with $\chi(z^\star)^\top a \neq 0$. Choose admissible nonzero g and set $\Delta_j = \epsilon a_j g$. Then $X_{\mathrm{train}}\Delta = 0$ while $\chi(z^\star)^\top \Delta = \epsilon(\chi(z^\star)^\top a) g$ is nonzero on a positive-measure set. The two models are training-equivalent but disagree at the target; Proposition A.2 gives positive SICR. Hence the row-span condition is necessary and sufficient. □

A.4.4 Interpretation of the quotient and span condition

Outcome-independent normalization and training-invisible interactions play different roles. The former is a gauge symmetry of one law; the latter can change the target law. Row-span coverage eliminates the target-relevant part of the latter ambiguity, even if the full interaction parameter remains unidentified.

B Structural Support Diagnostics

B.1 Exhaustive Support Enumeration

For each $K \in \{2, 3, 4\}$, $\mathcal{Z}_K = \{0, 1\}^K$. The residual interaction representation uses only $\chi(z) = (z_i z_j)_{i<j}$; atomic coordinates are excluded. For every target $z^\star$, every support $S_{\mathrm{train}} \subseteq \mathcal{Z}_K \setminus \{z^\star\}$ is enumerated. The pair is `IDENTIFIED_BY_INTERACTION_SPAN` iff $\chi(z^\star) \in \mathrm{rowspan}(X_{\mathrm{train}})$, and `AMBIGUOUS_BY_NULLSPACE` otherwise. Exact enumeration gives $K = 2 : 32/8/24$, $K = 3 : 1024/448/576$, $K = 4 : 524288/219104/305184$ (total/ambiguous/identified), grand total $525344/219560/305784$. Exact rank is primary with numerical cross-check; the archived coordinate-audit record reports 12,589,120 permutation checks with zero failures. The verification script reruns the exact enumeration and numerical rank checks and reads the archived permutation summary. The counts describe the specified binary pairwise family.

C Experimental Protocols and Additional Evidence

C.1 Finite-Sample Structural Plateau in a Stochastic Dynamical System

The finite-sample plateau experiment uses state $u_t = (x_t, v_t)^\top$, step size $dt = 0.05$, and 20 steps:

$$u_{t+1} = A(z)u_t + (0, dt\, F_s(z))^\top + (0, \sigma\sqrt{dt}\, \epsilon_t)^\top,$$
$$A(z) = \begin{pmatrix} 1 & dt \\ -dt(1 + 0.25 z_2 - 0.1 z_1 z_3) & 1 - dt(0.35 + 0.2 z_3) \end{pmatrix},$$

where $\epsilon_t \sim \mathcal{N}(0,1)$ independently and $\sigma = 0.12$. The learner is given z and the dynamics matrix $A(z)$ and estimates only the forcing coefficients from trajectory-mean velocity residuals. Its composition-dependent forcing is $F_s(z) = 0.3z_1 + 0.15z_1z_2 + s\,0.08z_2z_3$, for $s \in \{-1, +1\}$. The ambiguous support $S_{\mathrm{amb}} = \{000, 100, 010, 001, 110, 101\}$ never activates z_2z_3, whereas adding 011 supplies $\psi(011) = (0, 0, 1)$. Forcing fits use $\psi = (z_1, z_1z_2, z_2z_3)$ and the minimum-norm least-squares solution on composition-mean residuals.

Here n is the number of trajectories per composition, so the two supports use $6n$ and $7n$ trajectories; Appendix C.3 makes the budget-matched comparison. Target loss is the terminal-state Gaussian KL averaged over $(x_0, v_0) \in \{-1, -0.5, 0, 0.5, 1\}^2$. Writing $E = \widehat{F}(111) - F_s(111)$ for the target forcing estimation error, this loss is $L = \kappa E^2$, where $\kappa \approx 34.654424$ is the terminal Gaussian KL coefficient. The unestimated forcing coefficient therefore gives the analytic limit $\kappa \times 0.08^2 \approx 0.221788$.

For $n \in \{1, 2, 5, 10, 20, 50, 100, 200\}$, the plotted seed means for $s = +1$ on ambiguous support are 0.6383, 0.4447, 0.2667, 0.2504, 0.2303, 0.2247, 0.2203, and 0.2247. The evaluation uses 200 random seeds and contains $6{,}400$ seed-level metric records across both worlds and supports.

C.2 Accurate Marginal Prediction with an Unidentified Target Law

The latent composition Z is uniform on S_{amb} and hidden from the learner for Task M. For the trajectory-mean residual, the observable population law is the equal-weight Gaussian mixture

$$P_M^\star = \frac{1}{6}\sum_{j=1}^{6} \mathcal{N}(m_j, 0.0144), \qquad (m_1, \ldots, m_6) = (0, 0, 0, 0.3, 0.3, 0.45).$$

For each of 200 seeds and $n \in \{1, 2, 5, 10, 20, 50, 100, 200\}$, a six-component equal-weight mixture is fitted to an unlabeled prefix of residual trajectories; neither Z nor composition aggregates are exposed. Here n counts total unlabeled trajectories, whereas in Appendix C.1 it counts trajectories per composition. Regret for Task M, $D_{\mathrm{KL}}(P_M^\star \| q)$, is evaluated by deterministic quadrature. The worlds share the same training law and the same realized sample, so a single fitted marginal predictor gives their common loss; the experiment records store this loss for both worlds in 3,200 rows, with 1,600 quadrature reference records.

At $z^\star = 111$, however, the mechanism-resolved residual laws, after subtracting the common known target offset $0.45\mathbf{1}_{20}$, are $\mathcal{N}(+0.08\mathbf{1}_{20}, 0.288I_{20})$ and $\mathcal{N}(-0.08\mathbf{1}_{20}, 0.288I_{20})$. Their KL divergence is $\frac{1}{2}(0.16\mathbf{1}_{20})^\top (0.288I_{20})^{-1}(0.16\mathbf{1}_{20}) = 8/9$, so Task R remains structurally ambiguous. In contrast, mean regret for Task M falls from $R_M(1) = 2.167472$ to $R_M(200) = 0.009487$. This is the controlled separation: marginal prediction becomes accurate while the composition-conditioned target law remains unidentified.

C.3 Equal-Budget Support Intervention

The equal-budget intervention keeps $N_{\mathrm{total}} = 120 + M$ fixed while comparing same-span replication with targeted allocation to composition 011. The same-span interaction design remains rank 2, whereas targeted support has rank 3 whenever $M > 0$. The evaluated M grid is $\{0, 6, 12, 24, 48, 96, 192, 384\}$. At $M = 384$, the same-span mean loss is 0.228564, targeted mean loss is 0.026304, and their difference is 0.202260. The sampler uses 200 seeds per world. For cell j, let μ_j be the analytic mean loss and s_j the standard error of its 200-seed mean under the Gaussian forcing-estimation law $L = \kappa E^2$, with E and κ defined in Appendix C.1. The diagnostic $T_D = \max_j |\overline{L}_j - \mu_j|/s_j$ compares 32 cells against the two-sided Bonferroni normal reference $\Phi^{-1}(1 - 0.05/(2 \times 32))$. This descriptive check gives $T_D = 0.965827 < 3.162818$. These results support the claim that changing support can remove the structural plateau under the same total data budget.

C.4 Multi-Candidate Sequential Design

To avoid the unique-bridge geometry of the three-factor example, we construct 200 tasks from seeds 0–199 with six binary factors and the complete 15-dimensional pairwise map. For each task we

uniformly draw a target and eight initial non-target compositions, repeating only until the target residual is nonzero. Every remaining non-target, non-support composition is a candidate. The target itself is never selectable. Four label-free policies make six sequential choices: TSA; regularized D-optimal gain $\log(1 + \chi(c)^\top (X^\top X + 0.1I)^{-1}\chi(c))$; maximum candidate residual norm; and seeded random selection. Every policy starts from the identical task-specific support. For TSA, D-optimal, and maximum-residual selection, scores within 10^{-10} of the maximum are treated as tied; the smallest binary composition code is selected. The stored records include the selected composition at every step.

We measure the squared target residual after each choice, normalized by its initial value. At budgets one, two, and three, TSA means are 0.2931, 0.0362, and 0.0023; at budget three the D-optimal, maximum-residual, and random means are 0.5057, 0.5047, and 0.6257. TSA identifies 18.5%, 77.5%, 98.0%, and 100% of tasks after one through four choices. At budget six, D-optimal, maximum-residual, and random identify 9.0%, 10.0%, and 16.0% of tasks, respectively. Figure 4 reports means and 95% normal intervals across tasks. This benchmark tests only acquisition geometry: it does not use outcomes, sampling costs, or finite-sample response noise.

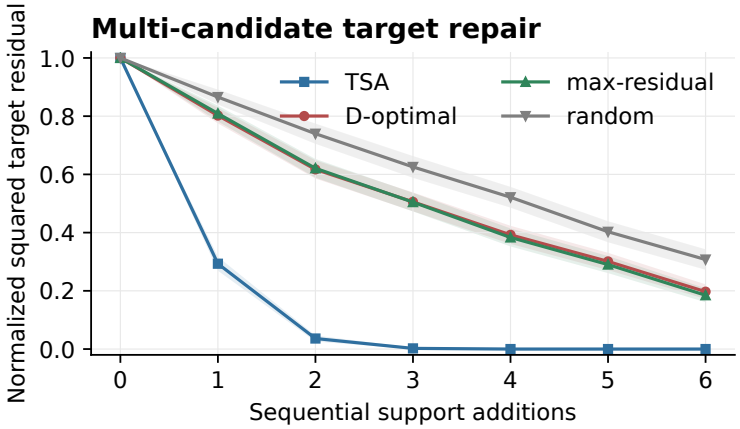


Figure 4: **Multi-candidate sequential support design.** Normalized squared target residual across 200 six-factor tasks. Unlike global-volume and residual-size baselines, TSA prioritizes the declared target direction. Bands are 95% normal intervals over tasks.

D Omitted-Interaction Stress-Test Protocol

Let $Z \in \{0,1\}^3$ and $Y \mid Z = z \sim \mathcal{N}(\mu(z), 1)$, where

$$\mu(z) = \alpha_0 + \sum_i \alpha_i z_i + \sum_{i<j} \beta_{ij} z_i z_j + \gamma z_1 z_2 z_3.$$

The assumed pairwise map is $\chi_2(z) = (z_1 z_2, z_1 z_3, z_2 z_3)$ and the enriched map is $\chi_3(z) = (\chi_2(z), z_1 z_2 z_3)$. Training uses every binary composition except 111, including the pairwise bridge 011. Hence $\chi_2(111)$ lies in the pairwise training span, but the cubic coordinate is zero on every training row and one at the target. A pairwise extrapolator misses the target mean by exactly γ; equal-variance Gaussian KL is therefore $\gamma^2/2$. Figure 5 evaluates $|\gamma| \in [0,1]$ on a 51-point grid. No Monte Carlo sampling is needed.

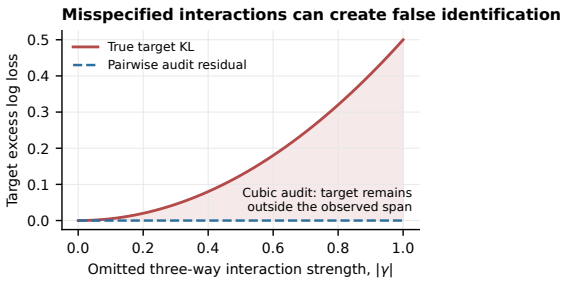


Figure 5: **Detecting an omitted interaction.** The pairwise audit reports zero residual after pairwise span repair, while the omitted three-way effect yields positive target KL. The enriched cubic audit leaves the target uncovered.

E Bike-Sharing Protocol and Additional Results

E.1 Data Provenance and Leakage Audit

We downloaded the CC BY 4.0 UCI Bike Sharing archive from the official repository.[1] The table has 17,379 rows and no missing values in the published fields. The SHA-256 hash of `hour.csv` is recorded below, followed by the exact leakage check:

`e03de4ee4ef4dc376ac6e04bf829673c6269e8eba5c60fa121640fa2f829504f`

$$\max_i |\texttt{cnt}_i - \texttt{casual}_i - \texttt{registered}_i| = 0.$$

Thus the two components reveal the target exactly and are excluded. Figure 6 contrasts the identity with leakage-safe random-split baselines. The audit uses the UCI source table; the Kaggle competition has a separate evaluation protocol.

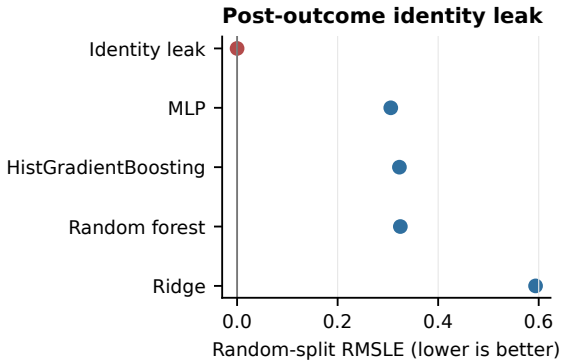


Figure 6: The post-outcome identity in the UCI source table obtains zero RMSLE, whereas leakage-safe random-split means across five seeds range from 0.306 to 0.594.

E.2 Features, Models, Splits, and Metrics

The leakage-safe input features are `season`, `yr`, `mnth`, `hr`, `holiday`, `weekday`, `workingday`, `weathersit`, `temp`, `atemp`, `hum`, `windspeed`, and integer elapsed day. Categorical variables are one-hot encoded; continuous variables are standardized. The response is $\log(1 + \texttt{cnt})$ and evaluation uses RMSLE after clipping count predictions at zero. Implementations use scikit-learn (Pedregosa et al., 2011): Ridge ($\alpha = 10$), random forest (120 trees, minimum leaf 2, feature fraction

[1] `https://archive.ics.uci.edu/static/public/275/bike+sharing+dataset.zip`

0.8), HistGradientBoosting (150 iterations, learning rate 0.06, at most 31 leaves, $\ell_2 = 0.2$), and MLP (two hidden layers of widths 64 and 32, early stopping, at most 220 iterations).

Table 1: Leakage-safe bike-demand RMSLE and efficiency. "Split" is a conventional 80/20 split and is not budget-matched. Gap is the unadapted support holdout; Same, Rand., Disagr., and TSA use equal acquisition budget $M = 200$. Timings are medians for TSA at $M = 200$ on one CPU thread.

Model	Split	Gap	Same	Rand.	Disagr.	TSA	Fit (s)	ms/1k
Ridge	0.594	0.859	0.801	0.798	0.792	0.680	0.012	8.85
Random forest	0.325	0.890	0.818	0.807	0.803	0.657	0.514	31.79
HistGradientBoosting	0.323	0.878	0.681	0.683	0.731	0.643	0.390	21.13
MLP	0.306	0.556	0.503	0.507	0.520	0.496	0.808	9.98

The six initial compositions are $\{000, 001, 010, 100, 101, 110\}$, with 100 uniformly sampled rows each. For seeds 0–9, nested budgets $M \in \{0, 25, 50, 100, 200, 400\}$ are selected by four rules that do not use candidate outcomes: (i) independent random permutations of unused same-span rows; (ii) a uniform random permutation of every unused non-target row (random-pool); (iii) descending cross-model predictive disagreement, computed from the four forecasters fitted only to the initial support (model-disagreement); and (iv) a random permutation of the 011 bridge pool (TSA). All 758 target-111 records form the fixed evaluation set and are never available to a selector. This retrospective support holdout uses non-target records from both calendar years, so it evaluates composition extrapolation rather than future-time forecasting. In this binary three-factor audit, 011 is the only non-target composition that raises interaction rank, so a rank-gain rule coincides with TSA; Appendix C.4 supplies the multi-candidate comparison where the objectives differ.

At $M = 200$, seed-wise averages across the four model classes are 0.700752 (same-span), 0.698598 (random-pool), 0.711623 (model-disagreement), and 0.619050 (TSA). Paired seed-wise differences relative to TSA are 0.081702 for same-span (95% normal interval $[0.060772, 0.102633]$) and 0.079548 for random-pool ($[0.056566, 0.102531]$); these descriptive intervals quantify variation over ten training/acquisition seeds for a fixed dataset and test set, after averaging the four models within each seed. The selected sets contain on average 0, 10.3, 2.9, and 200 bridge-011 rows, respectively, making the support mechanism directly auditable rather than inferred from the scores. We highlight $M = 200$ because it is shared with the cutoff study and report the full budget grid. At the largest budget, $M = 400$, the corresponding RMSLE averages are 0.656382, 0.649365, 0.689175, and 0.583044, while the bridge-row averages are 0, 19.7, 7.9, and 400. The paired random-pool–TSA difference is 0.066321 ($[0.044357, 0.088286]$). TSA therefore remains best on average, although MLP performs worse with TSA (0.476595) than with same-span acquisition (0.463723). Figure 7 reports the full budget trajectory for every model.

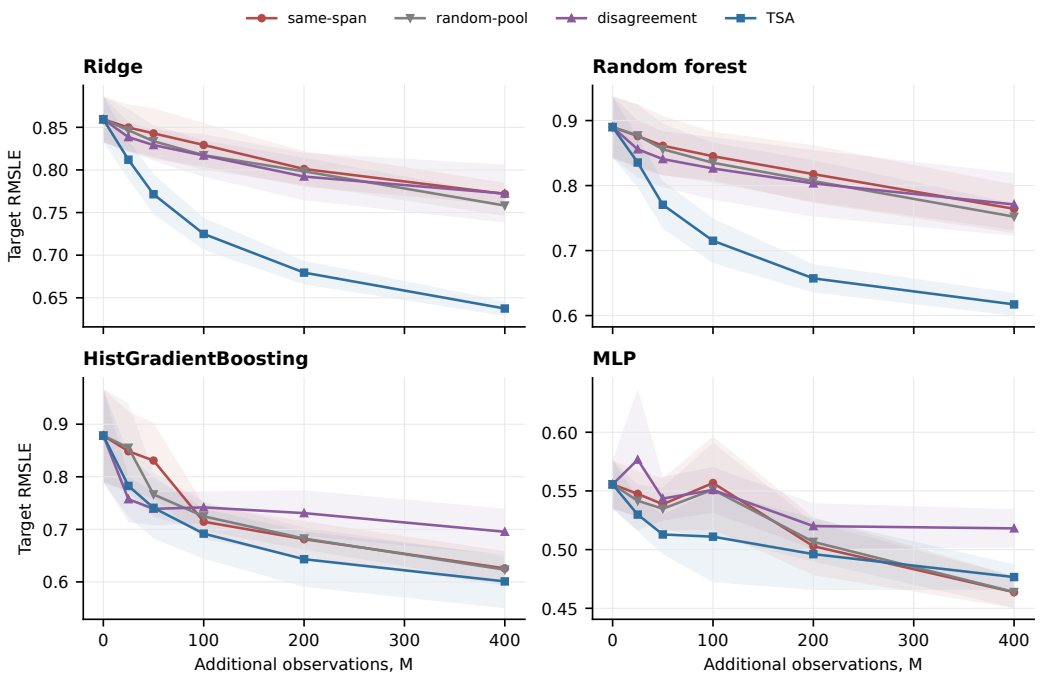


Figure 7: **Full equal-budget real-data results.** Target-111 RMSLE for each forecaster and acquisition policy over all evaluated budgets; bands are 95% normal intervals over ten seeds.

E.3 Cutoff Sensitivity and Timing

The sensitivity analysis uses three rush-hour definitions: tight (hours 8 and 17), standard (7–9 and 16–18), and broad (6–10 and 15–19), with 80 initial rows per composition, five seeds, HistGradientBoosting, and $M \in \{0, 25, 50, 100, 200\}$. At $M = 200$, TSA versus same-span mean RMSLE is 0.548 versus 0.588 (tight), 0.733 versus 0.818 (standard), and 0.750 versus 0.848 (broad). The ordering is stable, while the absolute difficulty changes with the target definition.

Fit and inference timings include preprocessing, use one thread, and are medians across the 10 TSA runs at $M = 200$ in the archived experiment. The support-audit timing repeats the $d = 3$, eight-candidate calculation 20,000 times inside each of nine trials and reports their median per-call time ($57.8\,\mu$s).