WHAT DOES AN OOD SPLIT MEASURE? DISTANCE, IDENTIFIABILITY, AND COMPLETION PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-distribution (OOD) benchmarks can be difficult for two different reasons: a test point may be far from the training inputs, or the training data may not contain enough information, under the chosen representation, to determine its prediction uniquely. We separate these cases on four dense six-component catalyst libraries. For each pair of elements (A, B), we hide every measured composition containing both elements and ask models to predict those held-out compositions from data in which A and B are observed separately but never together. A model-class-specific null-space audit shows that these test compositions can remain inside the training convex hull even though the A–B interaction is exactly unidentified; conversely, more distant composition-cap splits can remain fully identifiable. The resulting missing direction must therefore be filled by each estimator's inductive bias, or *completion prior*. An anchored difference transducer (ADT) recovers about 55–60% of the variance of the hidden interaction. Mechanism tests show that this recovery comes from a smooth separable model of property changes across composition space, rather than from low-rank structure in the pair-interaction matrix or lookup of previously observed moves. Exactly separable exponential-polynomial surfaces provide a positive control, while non-separable contamination rapidly removes the advantage. Finally, we test whether this structural match can be detected without using the true censored labels. A training-only "shadow censoring" diagnostic substantially lowers real-data model-selection regret relative to ordinary in-support holdout, although neither diagnostic reliably identifies the best method for each individual pair. The results distinguish geometric distance, design identifiability, and statistical predictability, and show that a formally unidentified interaction can still be predictable when the observed data contain transferable structure.

1 INTRODUCTION

Consider a compositional materials data set in which two elements, A and B, have each been measured many times but never together. A model is then asked to predict compositions that contain both. This is an OOD problem, but not necessarily because the test compositions are far from the training inputs. The harder issue may be that the training data never measured the A–B interaction at all. More generally, materials OOD benchmarks mix several kinds of novelty: distance from observed inputs, target values outside the observed range, and combinations of factors that were never jointly observed (Kauwe et al., 2020; Omee et al., 2024; Segal et al., 2025). These cases impose different statistical requirements, so treating them all as one notion of "OOD difficulty" can obscure what a model is actually being asked to infer.

Classical estimability theory gives a precise way to state the missing-information problem. For a model linear in parameters, $f_\theta(x) = \phi(x)^\top \theta$, the training data constrain only parameter combinations in the row space of the design matrix Φ_T; moving θ in a null-space direction $v \in \ker(\Phi_T)$ leaves every training prediction unchanged (Searle, 1987; Kounias and Chalikias, 2008). In a mixture model, for example, if A and B never co-occur, the feature $x_A x_B$ is zero on every training point. Its coefficient is therefore not identified by the training design, even if the held-out A–B compositions lie geometrically inside the observed composition cloud. We measure this pointwise with

$$\nu(x) = \frac{\|P_{\ker(\Phi_T)}\phi(x)\|_2^2}{\|\phi(x)\|_2^2}, \tag{1}$$

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.

where $\nu(x)$ is the fraction of the test feature vector lying in directions that the training design does not constrain. Throughout the paper, *identifiable* means that the chosen model class gives the same prediction for every parameter vector that fits the training data equally well. Thus $\nu = 0$ means the test prediction is fully determined under that feature map. Importantly, ν is a property of the *split together with the representation*; it is not a universal OOD score and is not, by itself, a predictor of error.

A separate line of work changes how extrapolation is represented. Pairwise Difference Regression learns property changes between observed examples and re-anchors those changes to measured references (Tynes et al., 2021). Bilinear Transduction instead represents a change as a product of a learned *displacement* representation and a learned representation of the starting composition, or *anchor* (Netanyahu et al., 2023); MatEx applies this idea to materials and molecules (Segal et al., 2025). This suggests a way an unidentified interaction might nevertheless be statistically predictable: responses to composition changes may share structure across different anchors and different element combinations. Classical results on finite-dimensional translation-invariant function spaces give an exact positive-control family for this idea: under standard regularity assumptions, functions with finite-dimensional translation structure are exponential polynomials (Anselone and Korevaar, 1964; Laczkovich, 2020). Recent OOD theory makes the complementary point that extrapolation is non-identifiable without some such structural commitment (Aguilar et al., 2026).

We study these questions on four dense experimental six-component catalyst libraries (Rohr et al., 2020), all measured on the same 0.1 composition grid. The grid lets us hide specific co-occurrences while keeping the underlying set of possible compositions fixed. This gives four linked questions. First, can geometric distance and design identifiability be varied independently? Second, when the training design no longer determines the answer, how do different estimators fill the missing interaction? Third, why can a bilinear transducer recover part of an interaction that was never observed directly? Fourth, can the suitability of that inductive bias be detected using only the available training data, before any truly censored label is revealed? We address the last question with *shadow censoring*: deliberately creating smaller missing-interaction problems inside the training set and asking whether a method succeeds on them.

2 Setup and Methods

The central prediction task is the following: remove all measured compositions containing a chosen element pair (A, B), fit a model to the remaining compositions, and predict the measured property on the hidden A–B compositions. The test measurements are used only for evaluation. The methods below separate three issues: how far the hidden compositions are from training data, whether the training design determines their predictions under a chosen model class, and what each estimator predicts when it does not.

Data and censoring tasks. Each library contains six elements and one measured catalyst property on compositions $x \in \mathbb{R}^6$, $x_i \geq 0$, $\sum_i x_i = 1$, in increments of 0.1, with at most four nonzero components. The complete grid has 2121 compositions; the four experimental libraries contain 2119–2121. One *fold* is defined by choosing a pair (A, B) and hiding every composition for which both $x_A > 0$ and $x_B > 0$. There are 15 such pairs per library and 60 folds in total. Training retains 69% of the compositions, so A and B are each still well represented individually; what is absent is their co-occurrence. In a quadratic mixture model this makes the interaction feature $x_A x_B$ exactly zero in training and nonzero on the test set. The prediction problem is therefore to infer the property of A–B mixtures without ever training on an A–B mixture.

We compare this main split with three controls. A *random holdout* measures ordinary interpolation. A *composition cap* removes compositions above a concentration threshold for one element; these points can be several grid steps from training while all polynomial interaction columns remain observed. Finally, a *censoring ladder* changes which interactions are missing. *Triple-co-occurrence censoring* removes only compositions containing A, B, and C together: pairwise AB, AC, and BC examples remain, so one cubic ABC interaction is lost but no quadratic interaction is. *Pair-co-occurrence censoring* is the main task: every AB co-occurrence is removed, eliminating one quadratic pair interaction together with higher-order terms that contain it. *Three-element isolation* is more severe: training keeps compositions containing at most one of A, B, C, so all three pairwise

interactions among the triple are missing. Triple censoring and pair censoring leave many nearby training compositions; three-element isolation also makes useful training anchors much sparser and therefore mixes algebraic censoring with a substantial increase in geometric distance. A *four-way co-occurrence control*, used in the appendix, removes only four-element co-occurrences and remains identifiable through cubic order. One grid substitution corresponds to moving 0.1 composition fraction from one element to another. Full masks, retained fractions, nearest-training distances, and anchor coverage are reported in Appendix B.

Identifiability audit. For each split we ask a model-class-specific question: if two parameter vectors make identical predictions on all training points, must they also make the same prediction at the test point? Polynomial feature maps of degree one to three define the design matrix Φ_T, and we compute Eq. (1) from its scaled SVD. A nonzero ν means that part of the test prediction lies in an exact training null space and therefore cannot be selected from the training labels alone. We measure geometric support separately using distance to the nearest training composition and, for the main censoring split, convex-hull membership. Ridge-weighted and RBF-kernel analogues are used in supporting analyses. Appendix A gives the estimability algebra and an explicit construction of two models that are identical on training data but differ on the censored test region.

Estimators and completion priors. When $x_A x_B$ is absent from training, a model must still assign some effect to the missing A–B interaction in order to predict the test set. We call the rule that supplies this unobserved component the model's *completion prior*. The central comparison uses four estimator families. Minimum-norm quadratic regression sets an exact null coefficient to zero. A cubic polynomial is fitted by least squares with a negligible stabilising ridge (10^{-6}) and is the reference for the RMSE contrasts below. An isotropic RBF Gaussian process (GP) continues the response according to kernel smoothness, with its length scale selected by ordinary in-support cross-validation. Finally, an *anchored difference transducer* (ADT) predicts property changes relative to nearby measured compositions.

For the transducer, an *anchor* a is an observed training composition near a query x_*, and the *displacement* is $\delta = x_* - a$. The learned change model is

$$h(\delta, a) = \sum_{k=1}^{r} f_k(\delta) g_k(a), \tag{2}$$

with neural networks $f, g : \mathbb{R}^6 \to \mathbb{R}^r$ and default latent width $r = 16$. ADT is trained on observed differences $\Delta y = y_b - y_a$. At prediction time, each nearby anchor gives $y_a + h(x_* - a, a)$, and ADT averages these anchored predictions over training points within three grid substitutions. The *absolute-target transducer* uses the same bilinear architecture but predicts the absolute target rather than a difference and does not add the measured anchor label. Because f takes the numerical displacement as input, ADT can evaluate a displacement that never occurred exactly in training. Additional difference models, structural priors, and the published MatEx implementation are defined in Appendix B.

Measuring what a model inferred about the missing interaction. Test RMSE tells us whether a predictor is accurate, but not whether it actually recovered the missing A–B interaction or merely fit the rest of the surface well. We therefore assign any predictor an *implied censored coefficient*. After producing predictions on the hidden A–B compositions, we combine those predictions with the observed training labels, refit the 22-column quadratic mixture model, and read off the fitted coefficient of $x_A x_B$. For mechanism analysis only, we compare this implied coefficient with the $x_A x_B$ coefficient fitted retrospectively to the complete experimental library. The true censored labels are never used to train the predictor. Recovery is summarized by correlation with that full-data coefficient and by R^2 relative to the baseline that fills the missing coefficient with zero. We also recompute ADT using only anchors whose displacement was seen during training or only anchors whose displacement was unseen, and compare with categorical move-table controls.

Synthetic tests of the transducer mechanism. The real-data result raises a specific mechanism question: what structure could allow ADT to infer an A–B interaction when no A–B composition appears in training? We test the hypothesis that property changes are approximately *separable* in displacement and anchor context. For a scalar response surface F, define its move field as

$D_F(a, \delta) = F(a + \delta) - F(a)$. The bilinear model in Eq. (2) assumes this field can be approximated by a small sum of products $f_k(\delta)g_k(a)$. Baseline synthetic surfaces instead use independent additive, pair, cubic, and smooth nuisance components on the same grid. As an exact positive control for separability we use

$$F(x) = \sum_{k=1}^{R} c_k \exp(2w_k^\top x), \qquad F(a+\delta) - F(a) = \sum_k [c_k e^{2w_k^\top a}][e^{2w_k^\top \delta} - 1], \tag{3}$$

so the move field has separation rank at most R. We sweep $R \in \{1, 2, 4\}$ and the transducer width. Denoting the separable surface in Eq. (3) by F_{sep}, we then mix it with an independent residual, $F_\alpha = \sqrt{1-\alpha}F_{\text{sep}} + \sqrt{\alpha}F_{\text{residual}}$, so increasing α progressively breaks the structure matched by the transducer.

Training-only shadow censoring. The mechanism experiments above use the true outer test labels for retrospective evaluation. To ask whether the same structural suitability could be detected *before* those labels are available, we create smaller missing-interaction tasks entirely inside the outer training set T_{AB}. For another pair (i, j) whose interaction is still observable in T_{AB}, we hide all ij co-occurrences, fit on $T_{AB,ij} = \{x \in T_{AB} : x_i x_j = 0\}$, and evaluate on $V_{ij} = \{x \in T_{AB} : x_i x_j > 0\}$. Because the ij interaction is estimable before this inner censoring, its reference value is known using only the outer training data. We use four such *shadow* pairs per outer fold, two sharing an element with A or B and two disjoint. The resulting coefficient-recovery and RMSE scores ask whether a model transfers correctly across deliberately missing interactions. A second score holds out complete displacement types to test generalization in δ. As a conventional baseline, ordinary validation randomly holds out 20% of the same outer training set, which tests interpolation rather than missing-interaction transfer. We compare the diagnostics retrospectively by how well they anticipate three outer-test outcomes: method-selection regret, the ADT–cubic RMSE contrast, and ADT's recovery of the censored coefficient. Here *regret* is the RMSE of the method selected using training data minus the RMSE of the best candidate method on the hidden outer test set; it is used only to evaluate the selector after the fact. Exact scores and bootstrap conventions are in Appendix B.

3 Results

3.1 Pair censoring can be geometrically easy but algebraically unidentified

We first ask what kind of OOD problem the main A–B censoring split creates. Geometrically, it is mild: across 12 folds spanning all four libraries, every examined censored test composition lies inside the convex hull of its training compositions. Algebraically, however, the split is incomplete. Under quadratic features the missing $x_A x_B$ column creates a nonzero null-space component, and mean ν rises from 0.009 for test points one grid substitution from training to 0.41 at distance five. Composition caps show the converse: they reach the same substitution distances while quadratic and cubic ν remain exactly zero. Random holdout also has $\nu = 0$. Thus distance from observed compositions and lack of information about a model parameter are empirically distinct axes (Fig. 1a).

Exact non-identifiability has a concrete consequence. For each censored fold we can construct two quadratic coefficient vectors that make essentially identical predictions on every training composition (largest median disagreement 8.5×10^{-17}) and have identical in-support CV error, yet their held-out RMSEs are 3.35σ and 3.22σ. Training performance therefore cannot distinguish these two continuations; selecting between them requires an assumption not determined by the observed labels. The audit also identifies *which order of interaction is missing*. Under triple-co-occurrence censoring, where pairwise interactions remain observed and only one cubic interaction disappears, the cubic model is not worse than the quadratic (0.0228 versus 0.0249 V). Identifiability is representation-dependent as well: on Materials Project data, censoring an element pair gives raw-fraction quadratic $\nu = 0.09$–0.11, but the same split has $\nu = 0$ in the 44-feature Oliynyk representation used by MatEx (Appendix B).

3.2 Model rankings depend on what information the censoring removes

We next use the ladder to ask how different completion priors behave as the training data lose progressively more interaction information. Triple-co-occurrence censoring removes a cubic interaction

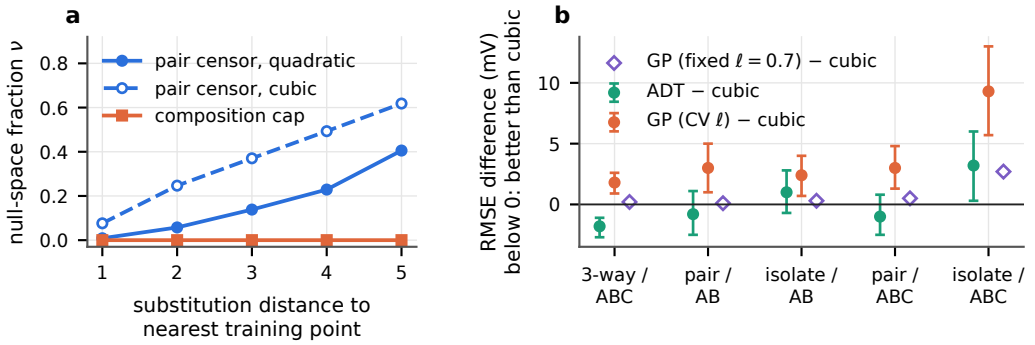

Figure 1: **(a)** Geometry and identifiability vary independently. Pair-censored test compositions are inside the training convex hull but require polynomial directions absent from training; composition-cap points reach the same nearest-training distances with $\nu = 0$ in both the quadratic and cubic bases. **(b)** The censoring ladder changes what the training set is missing. *3-way*, *pair*, and *isolate* are figure-only shorthand for triple-co-occurrence censoring, pair-co-occurrence censoring, and three-element isolation, respectively. "ABC" denotes test compositions containing all three selected elements; "AB" denotes compositions containing A and B but not C. Shown are RMSE differences from the cubic on the relevant test region; values below zero favour the plotted method. Error bars are 95% bootstrap intervals clustered by independent training set. Open diamonds show a GP with fixed $\ell = 0.7$, chosen after inspecting held-out error and therefore used only as a diagnostic.

while leaving all pairwise interactions observed; pair-co-occurrence censoring removes the AB pair interaction; three-element isolation removes three pair interactions and also pushes test points much farther from useful training anchors. Under the GP length scale chosen by ordinary in-support CV, GP–cubic RMSE grows from $+1.8\,\mathrm{mV}$ [$+0.9$, $+2.6$] under triple censoring in the ABC region to $+9.3\,\mathrm{mV}$ [$+5.7$, $+13.0$] under three-element isolation in the ABC region (Fig. 1b).

This trend is not simply evidence that an RBF smoothness prior fails whenever identifiability is lost. In-support CV selects a short length scale (0.25–0.35) in 57 of 60 real folds. Re-evaluating the ladder with a fixed, longer $\ell = 0.7$ removes almost all of the GP deficit on the four milder rungs. Only three-element isolation in the ABC region retains a clear deficit, $+2.7\,\mathrm{mV}$ [$+0.5$, $+4.9$], and three-element isolation is also where the mean nearest-anchor distance jumps from 1.28 to 3.51 substitutions and local anchor coverage collapses. The observed GP failure therefore combines three effects: the kernel's completion rule, a hyperparameter chosen only for in-support interpolation, and the increasing geometric gap created by the hardest censoring.

On the original 60 pair-censored folds, where every method is evaluated on the same 38,154 points, ADT has the lowest RMSE among the tested estimators (Table 1). Over five seeds, ADT also improves on the otherwise identical absolute-target transducer by $1.6\,\mathrm{mV}$ [1.1, 2.2], showing that re-anchoring the learned difference to the measured anchor label is useful in this regime. The released MatEx implementation performs substantially worse and is optimistically biased in all four libraries; because it differs from the absolute-target transducer in several implementation choices, we treat this comparison as a baseline rather than as evidence about bilinear transduction in general (Appendix B).

The transducer advantage is local rather than universal. ADT–cubic changes from $-1.8\,\mathrm{mV}$ [-2.7, -1.1] under triple censoring in the ABC region to $+3.2\,\mathrm{mV}$ [$+0.3$, $+6.0$] under three-element isolation in the ABC region. At its default anchor radius, ADT can predict only 51% of ABC-region points under three-element isolation because the remaining training set contains too few nearby anchors. Allowing anchors up to six substitutions away restores full coverage, but ADT then trails the cubic by $7.6\,\mathrm{mV}$ [1.4, 13.1] on the same 4608 points. The method works best when the missing interaction can be inferred from nearby, structurally related examples; once censoring removes both the interaction and those local analogues, its advantage reverses.

Table 1: Pair-censoring benchmark: 60 folds and 38,154 common test points, seed 0. RMSE is in volts of overpotential. A negative bias means the model predicts a better catalyst than was measured. The "completion prior" column describes how each method fills a component that is not determined by the training design.

Method	completion prior	bias (V)	RMSE (V)
ADT (difference target + anchor label)	smooth separable move field	$+0.0021$	**0.0255**
Absolute-target transducer	smooth separable move field	-0.0000	0.0272
Cubic	near zero	$+0.0003$	0.0279
Quadratic, minimum norm	exactly zero	$+0.0075$	0.0306
GP (CV length scale)	kernel-smooth	$+0.0020$	0.0319
Published MatEx BLT	learned, single anchor	-0.0126	0.0354

3.3 ADT RECOVERS THE MISSING A–B INTERACTION FROM A SMOOTH SEPARABLE MOVE FIELD

Why can ADT predict the censored region at all when no A–B composition was used for training? The implied-coefficient analysis separates recovery of the missing interaction from ordinary accuracy on the rest of the surface. Across the real folds, ADT's implied $x_A x_B$ coefficient correlates 0.67–0.72 with the coefficient fitted to the complete libraries in two independent runs and explains about 55–60% of its variance relative to zero completion. The direct cubic reaches 0.37 and the CV-selected GP -0.22 (Fig. 2a). An error decomposition gives the same conclusion: ADT's advantage over the cubic is $-3.7\,\mathrm{mV}$ along the missing-interaction component but only $-0.7\,\mathrm{mV}$ elsewhere. Thus ADT is not merely fitting the already identifiable part of the response surface more accurately; its predictions contain information about the interaction that was removed from training.

The controls show that this information is not obtained by simply reconstructing a low-rank matrix of pair coefficients or memorizing observed composition changes. Low-rank completion of the 15 observable pair interactions does not predict the missing pair well, and hand-designed relational summaries explain only part of ADT's recovery. More decisively, 29.6% of the anchor–move evaluations used by ADT on censored test points involve exact displacement vectors that never occur in training—these are the moves that increase both A and B. Nevertheless, predictions built only from unseen-displacement anchors recover the missing coefficient as well as those built from seen displacements (correlations 0.683 and 0.664). A context-free control that uses only a mean response for each observed move plus the anchor label reaches 0.367 correlation, while low-rank completion of the observed move-by-anchor table reaches 0.610. Together these results indicate that ADT uses both shared anchor–move structure and the smooth dependence of $f(\delta)$ on the numerical displacement, allowing it to generalize beyond a table of observed move types.

Exactly separable synthetic surfaces test whether this proposed structure is sufficient, rather than merely correlated with ADT's success on the real data. For the exponential-ridge family in Eq. (3), ADT recovers the censored coefficient with correlation 0.99–1.00 for true separation ranks 1, 2, and 4 and for both same- and mixed-sign directions, beating the cubic in 71 of 72 seeds. Increasing ADT's latent width stops improving coefficient recovery once the width is roughly sufficient for the generator's effective separation rank. On the real data, by contrast, width one already reaches 0.65 correlation with the censored coefficient, while larger widths mainly improve total RMSE. We therefore do not claim that the catalyst surface literally has rank one; rather, a very low-dimensional separable component appears sufficient to carry most of the recoverable information about the missing interaction.

The contamination experiment tests how specific this match is. As an independent non-separable residual is mixed into the exact separable surface, ADT's advantage disappears quickly: it beats the cubic only for $\alpha \leq 0.25$ and the GP only for $\alpha \leq 0.1$ (Fig. 2b,c). We use α only as a controlled mechanism axis, not as an estimate of a physical "separable fraction" in the real libraries. A direct exponential-polynomial model fitted to the real data points in nearly the same missing-coefficient direction (correlation 0.68 with truth and 0.675 with ADT) but has poor scale and overall RMSE. The evidence therefore supports smooth finite-separation-rank translation structure as an inductive bias used by ADT, not the stronger claim that the full catalyst response is itself an exponential-polynomial surface.

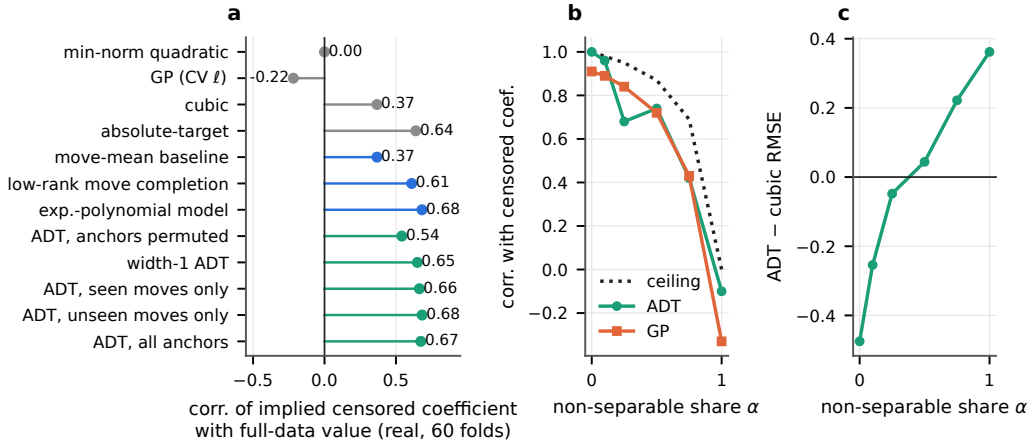

Figure 2: Evidence for the transducer mechanism. **(a)** Correlation between each method's implied missing $x_A x_B$ coefficient and the coefficient fitted to the complete real library (60 folds). Grey: standard estimators; blue: categorical and explicit-model controls; green: ADT variants. **(b)** On synthetic mixtures of separable and non-separable structure, coefficient recovery compared with the maximum correlation available from the separable component. **(c)** ADT–cubic RMSE on the same synthetic mixtures; values below zero favour ADT.

3.4 Shadow censoring tests this structural prior using training data only

The mechanism result is retrospective: it explains why ADT worked after the true A–B test labels are revealed. We next ask whether the available training data alone contain evidence that this completion prior is appropriate. Shadow censoring answers this by hiding additional interactions whose values are still knowable from the outer training set and measuring whether a method reconstructs them. Ordinary random holdout asks a different question: how well does the method interpolate among already observed kinds of compositions?

On the synthetic contamination worlds, where we know exactly how much non-separable structure was added, shadow scores track the intended mechanism. As α increases, width-one shadow R^2 falls from 0.80 to strongly negative values and the held-out-displacement score declines monotonically. The shadow ADT–cubic RMSE margin correlates $+0.70$ $[+0.48, +0.84]$ with the eventual outer-test margin, compared with $+0.44$ $[+0.19, +0.66]$ for the ordinary-holdout margin; the paired difference in these correlations is $+0.25$ $[+0.04, +0.50]$. However, when the diagnostics are used to choose among cubic, GP, and ADT, their mean regrets are statistically indistinguishable on these synthetic worlds (0.071 for shadow versus 0.092 for holdout; paired difference -0.021 $[-0.076, +0.036]$). Shadow therefore measures the separability axis more directly, but that does not automatically make it a better selector in every setting.

On the real catalyst folds, the training-only shadow score contains information about whether ADT will recover the hidden interaction. Width-one shadow R^2 correlates $+0.38$ $[+0.20, +0.58]$ with actual censored-coefficient recovery, and the held-out-displacement score gives $+0.24$ $[+0.07, +0.35]$. An ordinary-holdout proxy based on the cubic–ADT margin also correlates with recovery ($+0.28$ $[+0.01, +0.47]$), so the difference between those two recovery predictors is not established. The clearer practical difference appears when each diagnostic selects a model. The target-sharing shadow selector has mean regret 2.46 mV, compared with 5.46 mV for ordinary holdout; paired shadow-minus-holdout regret is -3.00 mV $[-6.55, -0.38]$. Shadow selects cubic/GP/ADT in 20/9/31 folds, close to the actual best-method counts 18/8/34, whereas ordinary holdout selects the GP in 40 of 60 folds. This reproduces the earlier problem: interpolation-based validation prefers the short-length-scale GP even though that choice transfers poorly to the missing-interaction task. Shadow selection still does not beat the simple fixed policy of always choosing ADT (mean regret 1.21 mV).

Shadow censoring is therefore a useful structural diagnostic but not a reliable pair-by-pair winner oracle. Neither shadow nor ordinary holdout predicts the exact real ADT–cubic margin reliably: their Spearman correlations are -0.12 $[-0.44, +0.11]$ and $+0.17$ $[-0.12, +0.25]$, respectively. The

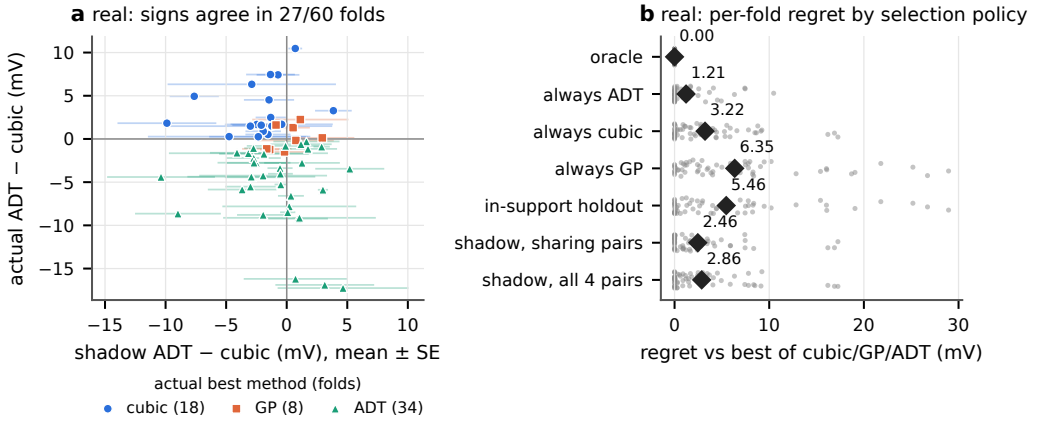

Figure 3: Training-only shadow censoring on the 60 real folds. **(a)** The shadow estimate of the ADT–cubic RMSE difference (mean ± SE over four inner censored pairs) versus the actual difference on the hidden outer A–B region. **(b)** Regret of model-selection policies relative to the best of cubic, GP, and ADT on each outer fold; diamonds show means. Selection uses no outer-test labels. The synthetic counterpart is in Appendix B.

differences between methods are themselves small: the median gap between the best and second-best method is $2.2\,\mathrm{mV}$, while the median standard error of the four-pair shadow ADT–cubic estimate is $2.4\,\mathrm{mV}$, comparable also to neural seed variation. The strongest supported conclusion is therefore narrower: training-only pseudo-censoring can reveal whether a model transfers across the same kind of missing interaction that defines the outer task, and on these real folds that structural match reduces model-selection regret relative to ordinary interpolation validation. It does not yet solve prospective method selection for an individual censored pair.

4 Discussion and Limitations

The experiments support a three-stage view of OOD prediction. *Geometric support* asks whether a query is near the observed inputs. *Design identifiability* asks whether the training set, under a chosen representation and model class, uniquely determines that query's prediction. *Statistical predictability* asks a different question again: if the prediction is formally unidentified, does the observed response surface contain regularities that let a particular inductive bias infer the missing component anyway? The pair-censoring experiments show that these questions can have different answers. The hidden A–B compositions are geometrically interior but their pair interaction is not identified, while ADT can still predict a substantial fraction of that interaction from structure elsewhere in the data.

This distinction limits how ν should be interpreted. The null-space fraction reports exact missing directions for a chosen feature map; it does not measure conditioning inside the identified row space, optimizer behavior, availability of nearby anchors, or whether a model's completion prior matches the data-generating structure. Correspondingly, ADT's success on the main pair-censoring task is conditional. Its advantage reverses when the censoring also removes nearby anchors, and it rapidly disappears when the synthetic move field is contaminated with structure outside its separable bias. The Levi–Civita positive control establishes that smooth finite-separation-rank structure is sufficient for recovery, not that the real catalyst surfaces are exactly exponential polynomials. The real-data evidence is also limited to four libraries, one measured property, and one composition grid; most neural results use a single seed, with a within-fold RMSE seed standard deviation of $1.5\,\mathrm{mV}$. The longer GP length scale used to diagnose the CV failure was chosen after inspecting held-out performance and is not a deployable hyperparameter choice.

Shadow censoring addresses a further practical issue: an explanation of OOD performance is much more useful if some evidence for it is visible before the OOD labels are exposed. In controlled worlds, the shadow scores follow the separability mechanism more closely than ordinary random holdout. On the real folds, shadow scores predict hidden-coefficient recovery and shadow-based selection has lower paired regret than ordinary holdout, mainly because it avoids holdout's strong preference

for the short-length-scale GP. These results support structurally matched pseudo-censoring as a way to probe whether an extrapolative prior is compatible with patterns already present in training data. They do not establish a general validation algorithm: shadow selection does not beat always choosing ADT on this benchmark, neither diagnostic reliably predicts the pair-specific ADT–cubic margin, and the recovery correlation is not significantly stronger than the best holdout proxy. More inner censoring tasks, repeated seeds, or explicit decomposition of pseudo-OOD error into identified and newly unidentified components may improve prospective selection, but that remains open.

The main conclusion is therefore not that one extrapolation method is universally best. It is that an OOD benchmark can ask three different questions—how far the test point is from observed inputs, whether the chosen model class is identified there, and whether an inductive bias can exploit structure to predict what the design leaves unspecified. Once the training design no longer determines a unique continuation, method comparisons partly compare these completion priors. In the catalyst data, transferable structure makes a formally missing interaction statistically predictable, and shadow censoring shows that some evidence about this compatibility is already present in the training data.

Reproducibility statement. Every reported number is traceable to a measurement register that maps it to the script and per-fold outputs that produced it; code, register and outputs will be released with the paper. Appendix B specifies the split construction, estimator settings, inference conventions, bootstrap clustering, synthetic generators, and supporting controls. The mathematical derivations underlying the identifiability and separability arguments are collected in Appendix A.

AI USE STATEMENT

Generative AI tools were used to assist with formulating and validating mathematical claims and writing mathematical proofs; implementing mathematical frameworks in code; developing and debugging code; running control experiments; providing early feedback on numerical accuracy; verifying numerical accuracy and internal numerical consistency; refining hypotheses; and supporting interpretation of experimental results. Generative AI tools also played a substantial role in writing and editing the manuscript, as well as in literature search and brainstorming. All AI-assisted mathematical claims, analyses, code, numerical results, and manuscript text were reviewed and verified by the authors, who take responsibility for the final content of the paper.

REFERENCES

L. Aguilar, J. Nagler, C. Hoelscher, and N. Antulov-Fantulin. Does your neural network extrapolate? Feature engineering as identifiability bias for OOD generalization. *arXiv:2605.07483*, 2026.

P. M. Anselone and J. Korevaar. Translation invariant subspaces of finite dimension. *Proceedings of the American Mathematical Society*, 15(5):747–752, 1964.

J. A. Cornell. *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, 3rd ed. Wiley, 2002.

S. K. Kauwe, J. Graser, R. Murdock, and T. D. Sparks. Can machine learning find extraordinary materials? *Computational Materials Science*, 174:109498, 2020.

S. Kounias and M. Chalikias. Estimability of parameters in a linear model and related characterizations. *Statistics & Probability Letters*, 78(15):2437–2439, 2008.

M. Laczkovich. The Levi-Civita equation in function classes. *Aequationes Mathematicae*, 94:689–701, 2020.

A. Netanyahu, A. Gupta, M. Simchowitz, K. Zhang, and P. Agrawal. Learning to extrapolate: A transductive approach. In *International Conference on Learning Representations*, 2023.

S. S. Omee, N. Fu, R. Dong, M. Hu, and J. Hu. Structure-based out-of-distribution (OOD) materials property prediction: a benchmark study. *npj Computational Materials*, 10:144, 2024.

B. Rohr, H. S. Stein, D. Guevarra, Y. Wang, J. A. Haber, M. Aykol, S. K. Suram, and J. M. Gregoire. Benchmarking the acceleration of materials discovery by sequential learning. *Chemical Science*, 11(10):2696–2706, 2020.

H. Scheffé. Experiments with mixtures. *Journal of the Royal Statistical Society: Series B*, 20(2):344–360, 1958.

S. R. Searle. *Linear Models for Unbalanced Data*. Wiley, 1987.

N. Segal, A. Netanyahu, K. P. Greenman, P. Agrawal, and R. Gómez-Bombarelli. Known Unknowns: Out-of-Distribution Property Prediction in Materials and Molecules. *npj Computational Materials*, 11:345, 2025.

M. Tynes, W. Gao, D. J. Burrill, E. R. Batista, D. Perez, P. Yang, and N. Lubbers. Pairwise Difference Regression: A machine learning meta-algorithm for improved prediction and uncertainty quantification in chemical search. *Journal of Chemical Information and Modeling*, 61(8):3846–3857, 2021.

A Mathematical foundations

A.1 Estimability and null-space demand

Let $\Phi_T \in \mathbb{R}^{n\times p}$ be the training design and $\phi_* \in \mathbb{R}^p$ a test feature vector. If $v \in \ker(\Phi_T)$, then θ and $\theta + v$ yield identical training predictions. The linear functional $\phi_*^\top \theta$ is estimable iff $\phi_* \perp \ker(\Phi_T)$, equivalently $\phi_* \in \mathrm{row}(\Phi_T)$. Writing P_N for the orthogonal projector onto the null space gives Eq. (1). We scale feature columns by their maximum magnitude over train and test before the SVD so an exactly censored column is not assigned an arbitrary scale. Rank uses threshold $s_k > 10^{-8} s_{\max}$.

A ridge-weighted analogue replaces the hard projector by spectral weights $\lambda/(s_k^2 + \lambda)$, and an RBF-kernel analogue is

$$\nu_{\mathrm{RKHS}}(x_*) = 1 - k_*^\top (K + \epsilon I)^{-1} k_*, \qquad k(x_*, x_*) = 1,$$

which is normalized noiseless GP posterior variance. These quantities depend on the feature map or kernel and therefore on the model class being audited.

A.2 Indistinguishable estimators and completion risk

Let N contain a basis of $\ker(\Phi_T)$ and choose the unit vector $v \in \mathrm{span}(N)$ that maximizes the test response norm. If $\hat{\theta}$ is a training least-squares solution, then

$$\theta_\pm = \hat{\theta} \pm cv, \qquad \Phi_T \theta_+ = \Phi_T \theta_-$$

for any c, while their test predictions differ by $2c\Phi_S v$. This makes exact non-identifiability operational.

For one censored coefficient A and any training-measurable completion $\widehat{A}(T)$, squared completion risk obeys

$$\mathbb{E}[(A - \widehat{A})^2] = \mathbb{E}[\mathrm{Var}(A \mid T)] + \mathbb{E}[(\mathbb{E}[A \mid T] - \widehat{A})^2].$$

If A is independent of the observed training information and mean zero, zero completion is Bayes-optimal. If A is conditionally predictable, the recoverable fraction is

$$\eta^2 = \frac{\mathrm{Var}(\mathbb{E}[A \mid T])}{\mathrm{Var}(A)}.$$

Thus formal unidentifiability does not imply statistical unpredictability; a prior can recover a null-space direction if the data-generating process couples it to observed structure.

A.3 Difference transduction and finite translation rank

For a scalar response surface F, define its move field

$$D_F(a, \delta) = F(a + \delta) - F(a).$$

The bilinear transducer assumes D_F is approximated by $\sum_{k=1}^{r} f_k(\delta) g_k(a)$. For any finite move-by-anchor evaluation table, the predicted matrix has rank at most r, but the stronger property is functional: f generalizes over the numerical displacement rather than treating each row as a category.

Finite-dimensional translation-invariant function spaces are exponential-polynomial under standard regularity assumptions (Anselone and Korevaar, 1964; Laczkovich, 2020). The generator in Eq. (3) is the simplest exact nonlinear subclass. More generally, exponential polynomials have schematic form $\sum_j P_j(x)e^{\lambda_j^\top x}$. We use pure exponential ridges as a positive control, not as a literal model of the catalyst surface.

A.4 Shadow censoring

For an outer censored pair (A, B), shadow censoring selects an estimable pair (i, j) inside T_{AB} and forms

$$T_{AB,ij} = \{x \in T_{AB} : x_i x_j = 0\}, \qquad V_{ij} = \{x \in T_{AB} : x_i x_j > 0\}.$$

The reference interaction L_{ij}^{ref} is fitted using all of T_{AB}; all fitting, early stopping and model selection in the shadow problem use only $T_{AB,ij}$. With implied completions $\hat{L}_{ij}$, we use

$$R^2_{\text{shadow}} = 1 - \frac{\sum_{ij}(\hat{L}_{ij} - L_{ij}^{\text{ref}})^2}{\sum_{ij}(L_{ij}^{\text{ref}})^2}.$$

Because the shadow split is nested inside an already censored outer fold, target-sharing and disjoint pseudo-pairs are reported separately. The score is a probe of structural transfer, not a claim that the inner and outer test distributions are identical.

B Experimental details and extended results

B.1 Data, noise, and split construction

The stored target is the negative overpotential; all reported real-data errors are converted to volts of overpotential. The integer grid coordinate is $z = 10x$, and the substitution distance is $\frac{1}{2}\sum_i |z_i - z'_i|$. Per-library noise scales estimated from fourth differences are 0.0199, 0.0141, 0.0156 and 0.0187 V. No composition was measured twice. The ladder uses 20 element triples per library. With presence indicators $p_i = \mathbf{1}\{x_i > 0\}$, pair-co-occurrence censoring keeps $\neg(p_A \wedge p_B)$, triple-co-occurrence censoring keeps $\neg(p_A \wedge p_B \wedge p_C)$, three-element isolation keeps $p_A + p_B + p_C \leq 1$, and the four-way co-occurrence control keeps $\neg(p_A \wedge p_B \wedge p_C \wedge p_D)$. Test points are split into an ***ABC region*** (A, B, C all present) and an ***AB-only region*** (A, B present, C absent). Table 2 gives the verified design; zero-column counts were re-derived from every saved fold. The pair-censoring rung covers 10 of the 15 pairs per library because only ordered triples are used.

Table 2: Censoring-ladder design. Nearest substitution distance and anchor counts are means over test points; coverage is the fraction of test points with at least one training anchor within three substitutions.

rung	zero quad. cols	zero cubic cols	training kept	nearest substitution distance	anchors ≤ 3	coverage
four-way control	0	0	96%	1.12	183.9	100%
triple censoring	0	1	86%	1.28	138.5	100%
pair censoring	1	7	69%	1.60	97.2	97.9%
isolation	3	19	34%	3.51	14.5	51.0%

Across 12 censored-pair folds covering all four libraries, convex-hull membership of every test point was verified by linear feasibility (hull fraction 1.00). The scarcest single element still appears in 587 training compositions of a censored fold. Mean quadratic ν of censored-pair points is 0.0086, 0.0576, 0.1384, 0.2289 and 0.4055 at nearest substitution distance 1–5 (cubic: 0.0762, 0.2467, 0.3704, 0.4932, 0.6182); caps and random holdout give exactly zero. The RBF-kernel analogue at $\ell = 0.12$ is 0.944 on caps and 0.574 on censored pairs, so kernel demand does not follow polynomial demand.

B.2 Estimator definitions and hyperparameters

The quadratic baseline uses the 22-column canonical mixture basis $(1, x, \{x_i x_j\}_{i<j})$ and minimum-norm least squares with no penalty. The cubic uses the 78-column degree-three audit basis with

combined-maximum column scaling and ridge 10^{-6} as a numerical stabilizer. The hierarchical interaction-prior baseline minimizes

$$\|\Psi_T\beta - y_T\|^2 + \tau^{-2}\sum_{i<j}(L_{ij} - \lambda_i - \lambda_j)^2,$$

with $\tau \in \{0.003, 0.01, 0.03, 0.1, 0.3\}$ by four-fold training CV. The GP standardizes the target and uses an isotropic RBF kernel; $\ell \in \{0.08, 0.12, 0.18, 0.25, 0.35, 0.5\}$ is chosen by four-fold in-support CV with dimensionless noise $\max\{(\sigma/\mathrm{sd}(y_T))^2, 10^{-4}\}$. Its posterior mean is kernel ridge regression with that penalty. A kernel-ridge variant chooses ℓ and the penalty jointly by CV from $\{10^{-4}, 10^{-3}, 10^{-2}, 0.03, 0.1, 0.3\}$. Fixed length scales $\ell \in \{0.7, 1, 3\}$ were evaluated as diagnostics only.

Training moves for difference methods are sampled uniformly from training-point pairs and restricted to substitution distance ≤ 3. Moves are drawn by sampling 300 000–400 000 index pairs with replacement and keeping those with positive substitution distance at most three. The absolute-target transducer and ADT use networks $6 \to 64 \to 64 \to r$ with SiLU activations, Huber loss threshold σ, Adam learning rate 3×10^{-3}, a 15% pair validation split, at most 60 000 pairs per epoch in batches of 4096, evaluation every five epochs, patience six evaluations, a 250-epoch cap, and default width 16. Test points without an anchor within three substitutions are abstained on, and coverage is reported. The published MatEx BLT code is evaluated unchanged on the six composition fractions: two ReLU networks $6 \to 256^{\times 4} \to 42$ combined by an inner product, absolute target, L1 loss, 8000 epochs, Adam at 10^{-3}, and a single anchor chosen by matching stored training displacements, with no anchor label and no averaging. Its saved predictions are rescored as RMSE. Additional baselines include the exact quadratic midpoint-difference model, a cubic extension, an exact-displacement lookup model, the absolute-target transducer, and kernel ridge variants.

B.3 Full catalyst benchmark and ladder controls

Table 3: Full pair-censoring benchmark, 60 folds, identical 38,154 test points, seed 0, volts of overpotential. The last column is the paired bootstrap RMSE difference from ADT.

method	bias	RMSE	MAE	corr.	RMSE − ADT
ADT	+0.0021	**0.0255**	0.0184	0.875	–
Absolute-target transducer	−0.0000	0.0272	0.0199	0.856	+0.0017 [+0.0015, +0.0019]
cubic	+0.0003	0.0279	0.0203	0.847	+0.0025 [+0.0023, +0.0027]
midpoint difference	+0.0036	0.0284	0.0207	0.844	+0.0030 [+0.0028, +0.0031]
quadratic (min. norm)	+0.0075	0.0306	0.0226	0.825	+0.0051
hierarchical interaction prior	+0.0019	0.0309	0.0229	0.814	+0.0054
GP (CV ℓ)	+0.0020	0.0319	0.0225	0.796	+0.0064
published MatEx BLT	−0.0126	0.0354	0.0246	0.776	+0.0099 [+0.0096, +0.0103]

Five-seed paired reruns give ADT minus the absolute-target transducer as $-1.6\,\mathrm{mV}$ [−2.2, −1.1], with ADT better in 71% of 300 fold–seed pairs; the within-fold seed standard deviation is 1.5 mV. The published BLT is optimistic in every library (−13.4, −9.3, −15.1, $-13.4\,\mathrm{mV}$ for Libraries 1–4). On identical catalyst folds, the released code scores 1.995 (mean over folds of RMSE/σ) against 1.520 for a local reimplementation of the same absolute-target method. Tested explanations (anchor distance, tie-breaking, loss, skew ordering, epoch count, architecture) account for at most 12% of this gap. Figure 4 shows the error distributions by library and Fig. 5 predicted against measured values on Library 1.

Stratified by the exact demand $x_A x_B$, GP−quadratic is $-2.5\,\mathrm{mV}$ [−3.2, −1.9] in the lowest tercile and $+3.8\,\mathrm{mV}$ [+0.1, +7.5] in the highest. ADT−cubic is $-2.1\,\mathrm{mV}$ [−2.9, −1.3] and $-2.7\,\mathrm{mV}$ [−5.2, −0.2], so ADT leads at every demand level on this split. On the 60 folds, GP RMSE is 0.0319 V with CV-selected ℓ, 0.0301 V as kernel ridge, and 0.0283 V at $\ell = 0.7$, against 0.0279 V for the cubic.

The full-pool ladder RMSEs are:

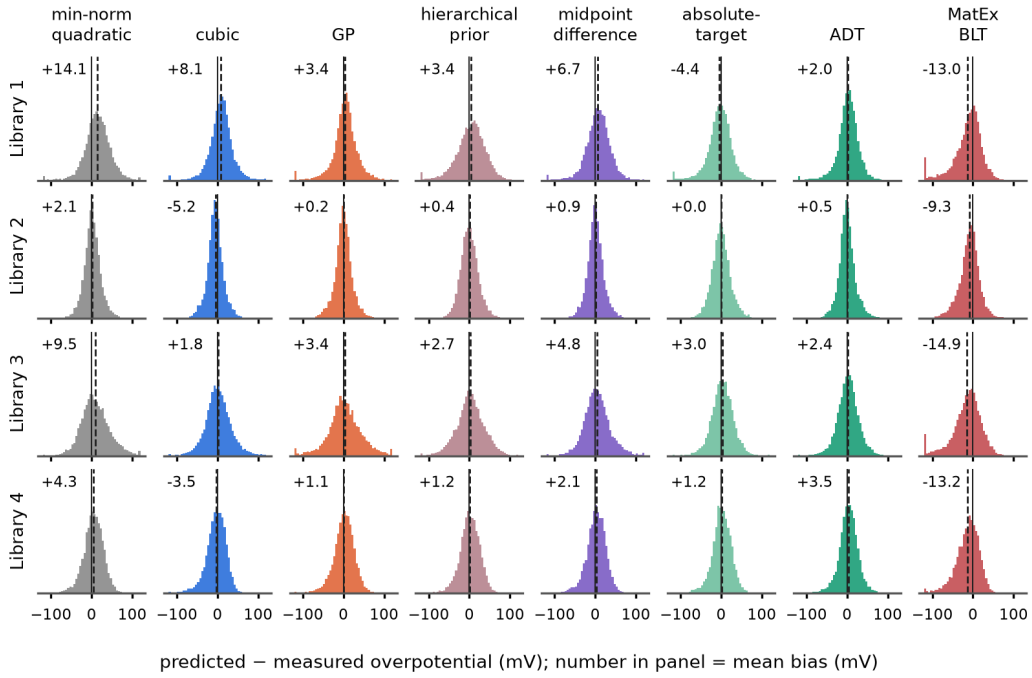

Figure 4: Held-out error (predicted minus measured overpotential, mV) on the censored-pair folds by library (rows) and method (columns), matched points. Dashed: mean; solid: zero.

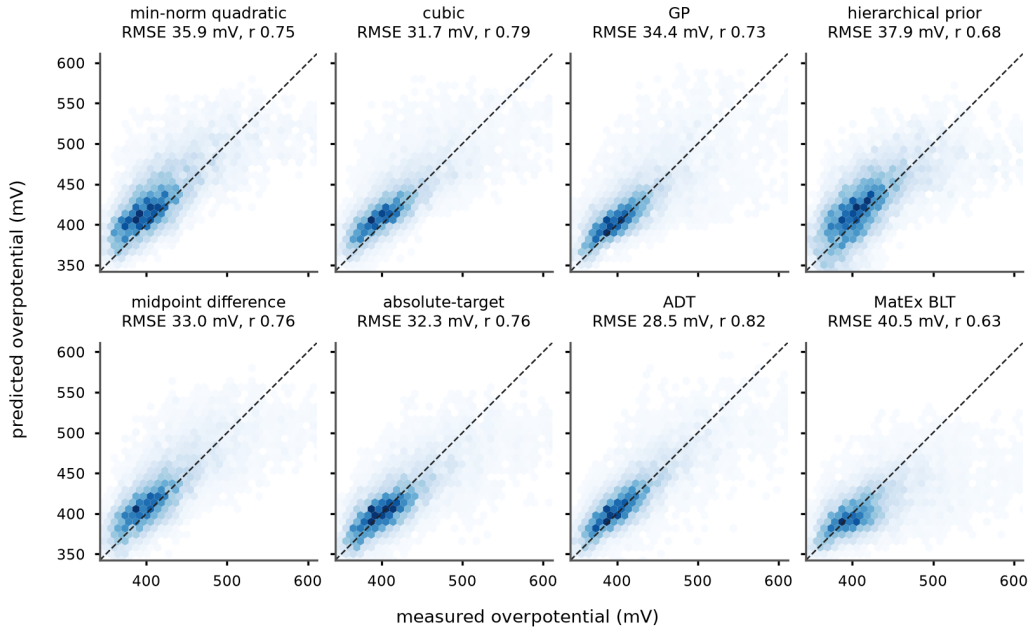

Figure 5: Library 1 pair-censoring folds, identical points. Top: error distributions. Bottom: predicted against measured overpotential on common axes. “Transducer” is the anchored difference transducer (ADT).

	triple censoring / ABC	pair censoring / AB	isolation / AB	pair censoring / ABC	isolation / ABC
Cubic	.0228	.0285	.0288	.0261	.0341
GP	.0245	.0311	.0309	.0289	.0413
ADT	.0209	.0277	.0298	.0251	.0373

The corresponding ADT−cubic contrasts are -1.8, -0.8, $+1.0$, -1.0 and $+3.2\,\mathrm{mV}$; only the endpoints exclude zero. Bootstrap intervals cluster folds that share a training set (40 clusters on pair censoring, 80 elsewhere). At a matched training size of 700 the endpoint contrasts are unchanged (-1.8 and $+2.9\,\mathrm{mV}$). With the GP at fixed $\ell = 0.7$ (seed 0), GP−cubic is $+0.2$, $+0.1$, $+0.3$, $+0.5$ and $+2.7\,\mathrm{mV}$ on the five rungs, and only the last interval excludes zero.

The ABC region under three-element isolation has mean nearest substitution distance 3.51, only 14.5 anchors within three substitutions, and 51% ADT coverage. On Library 1 and Library 3 (8 triples), refitting with anchor radius 4, 5 and 6 raises coverage to 80.2%, 96.9% and 100%. At radius six, ADT scores 0.0476 V against 0.0400 V for the cubic on the same 4608 points, a deficit of $+7.6\,\mathrm{mV}$ [$+1.4$, $+13.1$] clustered by triple. The newly covered points are predicted worse than those already covered within radius three.

B.4 Mechanism controls

The implied-coefficient instrument is calibrated by adding known $c\,x_A x_B$ perturbations; the measured slope is 1.000 and the training-only minimum-norm fit returns zero. The original ADT run gives correlation 0.719 with the full-data interaction and $R^2 = 0.60$ relative to zero completion; the later per-anchor refit gives 0.674 and 0.549. ADT's advantage over the cubic decomposes to approximately $-3.7\,\mathrm{mV}$ in the unidentified direction and $-0.7\,\mathrm{mV}$ elsewhere.

Five training-only relational summaries (additive completion, common-neighbour terms and response-signature similarity/distance) jointly explain 35% of the true coefficient and 47% of ADT's completion under leave-one-library-out prediction; ADT retains partial correlation 0.49 with truth after controlling them. Removing direct $A \leftrightarrow C$ and $B \leftrightarrow C$ bridge moves (about 4% of moves) changes RMSE by $-0.10\,\mathrm{mV}$ [-0.44, $+0.24$] relative to matched random removal.

Pure rank-1–4 completion of the 15-edge pair interaction graph gives correlations of -0.11, -0.36, -0.39 and -0.40 with the held-out truth; additive-plus-low-rank variants reach only 0.29–0.33. The residual move matrix after a quadratic fit is only slightly more low-rank than a quadratic-plus-noise surrogate: its top-three singular-value share exceeds the surrogate by 0.012–0.051. In contrast, seen- and unseen-displacement ADT anchors give essentially identical coefficient recovery. Categorical low-rank move completion reaches 0.610 correlation, while context permutation reduces ADT to 0.541. These controls motivate the smooth separable move-field interpretation rather than categorical matrix completion.

B.5 Synthetic generators, exactly separable surfaces and width

Generators. All synthetic surfaces live on the 2121-point grid. The mechanism experiments use one randomly chosen pair-censoring fold, 600 training points drawn from the uncensored compositions, and observation noise 0.15 in units of the clean-signal standard deviation. The independent-coefficient generator is a sum of standardised additive, pair, cubic and random-Fourier nuisance components (24 features, length scale 0.35) with independent standard-normal coefficients; the default shares are $(0.55, 0.30, 0.10, 0.05)$. With independent pair coefficients, zero completion is Bayes-optimal; in the mechanism run (shares $(0, 0.70, 0.25, 0.05)$, 12 seeds) ADT's completion has correlation -0.18 with the truth and excess risk $+0.70\,\mathbb{E}[A^2]$ relative to zero. The Levi–Civita generator draws $w_k \sim \mathcal{N}(0, I_6)$ (or $|w_k|$ for same-sign directions) and $c_k \sim \mathcal{N}(0,1)$, uses $\beta = 2$, and standardises F over the grid. The contamination residual F_{res} is the independent generator with shares $(0, 0.70, 0.25, 0.05)$, standardised; the ceiling at each α is the correlation across seeds between the full censored coefficient and its Levi–Civita part. The positive control uses 12 seeds for each of 3 ranks and 2 sign patterns; the width sweep uses 6 seeds per rank, and the contamination sweep 8 seeds per α.

For $F(x) = \sum_{k=1}^{R} c_k \exp(2 w_k^\top x)$, ADT coefficient correlation is 0.99–1.00 for $R = 1, 2, 4$ and both mixed- and same-sign directions; it beats the cubic in 71 of 72 seeds. Width sweeps show saturation once width is roughly sufficient for the generator's effective rank. On real data the width sweep gives coefficient correlations 0.650, 0.593, 0.649, 0.694, 0.652, 0.641 and 0.766 for widths 1–64, with single-seed variation comparable to known seed noise; RMSE improves from 0.0279 V at width 1 to about 0.0254–0.0257 at widths 8–32. The safe conclusion is that one latent factor already carries

most of the recoverable hidden-interaction signal, not that the full response surface has rank one (Fig. 6).

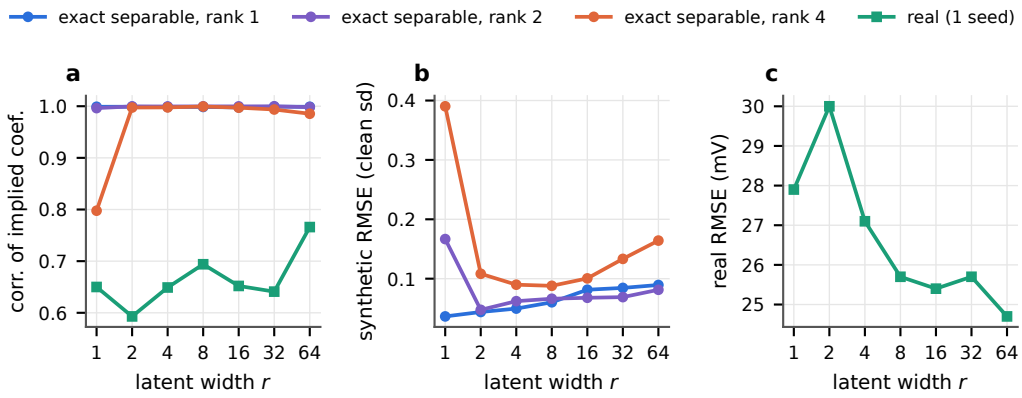

Figure 6: Width calibration. **(a)** Correlation of the implied censored coefficient with the truth against transducer latent width, for exact Levi–Civita surfaces of rank 1, 2 and 4 (6 seeds each) and for the real catalyst folds (single seed; seed standard deviation 0.047). **(b)** Synthetic RMSE. **(c)** Real pooled RMSE.

In the contamination family F_α with $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 1\}$, ADT's correlation with the censored coefficient is 1.00, 0.96, 0.68, 0.74, 0.42 and -0.10, against a ceiling of 1.00, 0.98, 0.95, 0.87, 0.69 and 0; the GP gives 0.91, 0.89, 0.84, 0.72, 0.43 and -0.33. ADT$-$cubic RMSE is -0.475, -0.254, -0.048, $+0.044$, $+0.222$ and $+0.362$. The direct exponential-polynomial model $b + v^\top x + \sum_k c_k e^{w_k^\top x}$ gives coefficient correlation 0.68 with truth and 0.675 with ADT but negative R^2 relative to zero completion and RMSE 0.0324 V, indicating correct completion direction but poor calibration.

B.6 Shadow censoring details

Four inner shadow pairs are used per outer real fold: two share an element with the censored pair and two are disjoint. The held-out-displacement score holds out entire displacement types from move-model fitting and compares the width-1 ADT with a context-free displacement baseline, implemented by fixing the anchor input: $S_{\text{sep}} = 1 - \text{MSE}(\text{ADT}_{r=1})/\text{MSE}(\text{context-free})$ on the held-out displacements. The shadow selector chooses, per fold, the method among cubic, GP and ADT with the lowest mean pseudo-held-out RMSE; the comparator is a random 20% holdout inside T_{AB}.

On synthetic contamination worlds, width-1 shadow R^2 is 0.80, 0.47, 0.07, -1.06, -2.47 and -2.35 across the six α values, and S_{sep} is 0.61, 0.48, 0.37, 0.27, 0.22 and 0.21. For predicting the actual ADT$-$cubic contrast, the shadow RMSE margin has Spearman correlation $+0.696$ [$+0.48$, $+0.84$] and sign agreement in 34/48 worlds, versus $+0.442$ [$+0.19$, $+0.66$] and 32/48 for ordinary holdout; their paired correlation difference is $+0.254$ [$+0.04$, $+0.50$]. Width-16 shadow R^2, width-1 shadow R^2 and S_{sep} have correlations -0.72 [-0.86, -0.47], -0.52 [-0.76, -0.31] and -0.45 [-0.73, -0.20] with the same contrast. Selector regret is 0.071 for all-four-pair shadow selection and 0.092 for in-support holdout; the paired difference is -0.021 [-0.076, $+0.036$], so no selection advantage is established on the synthetic worlds. Mean actual RMSE (clean units) is 0.311 for the oracle, 0.379 always GP, 0.383 shadow selection, 0.404 in-support holdout, 0.453 always cubic, and 0.472 always ADT (Fig. 7).

On real data, width-1 shadow R^2 correlates $+0.378$ [$+0.20$, $+0.58$] with actual coefficient recovery and S_{sep} gives $+0.242$ [$+0.07$, $+0.35$], clustered by library. Two ordinary-holdout proxies give $+0.282$ [$+0.01$, $+0.47$] for the cubic-minus-ADT RMSE margin and -0.287 [-0.47, -0.03] for negative ADT RMSE. The paired correlation difference between width-1 shadow R^2 and the holdout margin is $+0.097$ [-0.27, $+0.48$], so superiority over that proxy is not established; against negative holdout ADT RMSE the difference is $+0.665$ [$+0.25$, $+0.96$]. Neither diagnostic predicts the exact

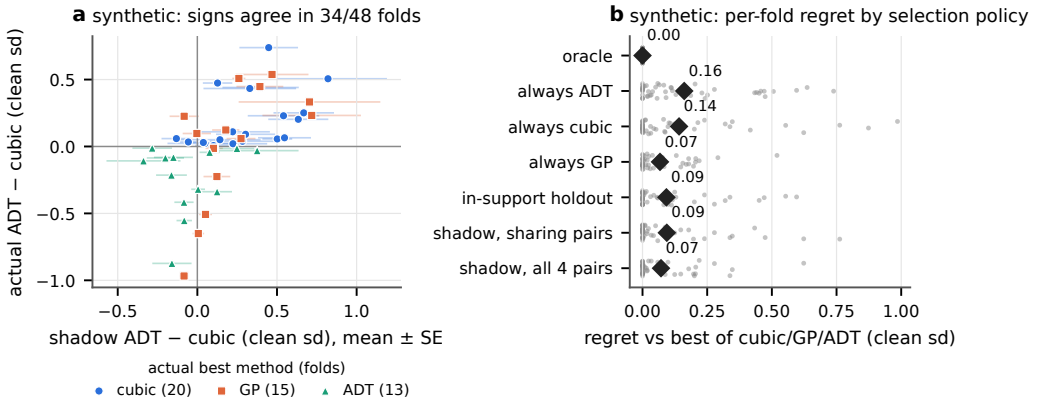

Figure 7: Shadow censoring on the synthetic contamination worlds (48 folds, clean-signal units), laid out as Fig. 3.

real ADT−cubic contrast reliably: shadow margin Spearman is -0.119 $[-0.44, +0.11]$ with 27/60 sign agreement, while holdout is $+0.166$ $[-0.12, +0.25]$ with 37/60.

Real per-fold method-selection RMSE is 0.0234 for the hindsight oracle, 0.0246 for always ADT, 0.0258 for target-sharing shadow selection, 0.0262 for all-pair shadow selection, 0.0266 for always cubic, 0.0288 for ordinary in-support holdout and 0.0297 for always GP. Relative to the per-fold oracle, mean regret is 2.456 mV for target-sharing shadow selection and 5.458 mV for holdout. Their paired difference is $-3.003\,\mathrm{mV}$ $[-6.547, -0.381]$: shadow is better on 32 folds, holdout on 16, with 12 ties. The all-four-pair shadow selector also improves on holdout, $-2.594\,\mathrm{mV}$ $[-5.818, -0.045]$. Target-sharing shadow selects cubic/GP/ADT in 20/9/31 folds, compared with actual-best counts 18/8/34; holdout selects 5/40/15. By library ($n = 4$, descriptive), width-1 shadow R^2 / S_{sep} / actual ADT−cubic (mV) is 0.34 / 0.47 / -4.1 for Library 3, 0.48 / 0.21 / -3.3 for Library 1, 0.06 / 0.14 / -0.5 for Library 4, and -2.79 / 0.07 / 0.0 for Library 2. Of the 60 real folds, the best of the three methods is ADT in 34, the cubic in 18 and the GP in 8. The median true best-versus-second-best margin is 2.2 mV [0.8, 4.4]; the median standard error of the four-shadow-pair ADT–cubic contrast is 2.4 mV. Only 7/60 folds have unanimous shadow-pair signs. Together, these results support using shadow censoring as a training-only probe of structural transfer rather than as a general per-fold winner oracle: it is strongly aligned with the controlled separability axis, reduces real selection regret relative to ordinary holdout, but does not reliably predict the exact real method margin.

B.7 Noise, averaging and uncertainty

Noise injection shows that anchor averaging is not primarily measurement-noise cancellation. Adding Gaussian noise of 0, 0.5, 1 and 2σ to the targets gives averaging gains (single anchor minus all anchors) of 8.4, 9.3, 12.1 and 22.0 mV; extrapolating to zero label noise leaves 5.1 mV, 61% of the gain at measured noise. The anchor-label control variate behaves oppositely: ADT minus absolute-target-transducer is -2.0, -2.2, -1.4 and $-0.3\,\mathrm{mV}$ across the same levels. Noise-corrected anchor-error correlation increases from 0.27 to 0.36 to 0.51 across increasing $x_A x_B$ demand, indicating an increasingly shared completion error that averaging cannot remove.

Nominal 90% intervals are close to nominal when pooled but differ after stratification by censored-interaction demand. With measurement variance added, low-to-high demand coverage changes from 93.4% to 84.5% for pseudo-inverse least squares, 95.2% to 86.7% for midpoint anchor dispersion, 96.1% to 85.6% for ADT dispersion, and 93.9% to 90.5% for the GP. Only the GP substantially widens its interval with demand, by 47% (Fig. 8). Pooled coverage is 89.7%, 92.2%, 91.6% and 91.5% for least squares, GP, midpoint and ADT.

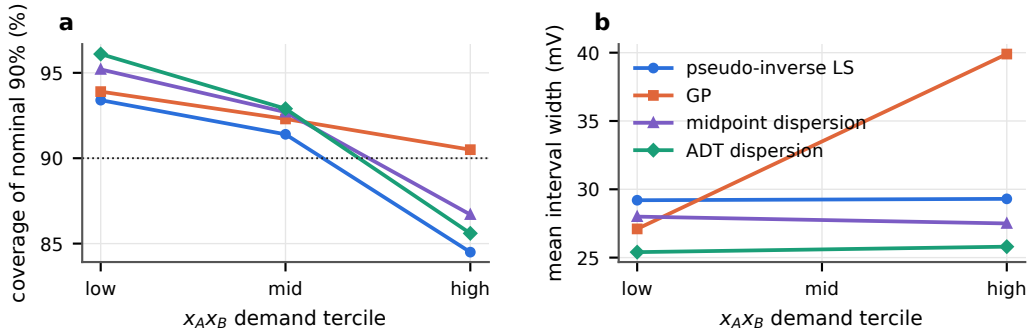

Figure 8: Nominal 90% intervals on the 60 real censored folds, measurement variance added in quadrature. **(a)** Coverage by tercile of the censored-column demand $x_A x_B$. **(b)** Mean interval width in the lowest and highest terciles.

B.8 Public benchmarks and evaluation controls

On continuous concrete data, exact null directions appear only in censoring tied to genuine structural zeros (quadratic $\nu = 0.152$ on 3 splits); all 143 continuous quadrant and cap splits have $\nu = 0$. In-support CV there occasionally selects a divergent ridge penalty (costs of $64\times$ and $34\times$ on 2 of 11 splits), and one of those splits has $\nu = 0$ exactly: ν does not predict numerical fragility (correlation with log cost at most 0.37).

On the MatEx data, co-occurrence censoring is infeasible on AFLOW (no element pair co-occurs in 50 or more compounds). On Materials Project bulk modulus, Table 4 shows that element-pair censoring is non-identifiable in raw fractions but fully identifiable in the 44-dimensional Oliynyk representation used by MatEx, while the published target-out-of-range split has the largest RBF-kernel demand.

Table 4: Null-space fraction of MP bulk-modulus test regions in three representations.

split	n_{test}	quadratic (raw fractions)	Oliynyk linear	RBF kernel
published target-out-of-range	315	0.000	0.000	0.286
random holdout	315	0.000	0.000	0.118
censor Li–O / Fe–O / Cu–O / Na–O	94–117	0.087–0.108	0.000	0.013–0.178

On MatEx's own tasks, where only the target is out of range, the released bilinear-transduction code reproduces the published MAE on all five AFLOW tasks attempted but not on MP bulk modulus (58.5 against 45.8). The published ridge baseline cannot be checked because no ridge code is released; a tuned ridge on the released features removes a median 62% of the published ridge-to-transduction gap. Table 5 compares six methods grouped by whether they add the anchor's measured label.

Table 5: Six methods on five AFLOW MatEx tasks, test MAE; best in each group in bold. The bilinear-ridge columns use an affine anchor map and are solved by ridge; the left variant predicts the absolute target, while the right predicts a difference and adds the anchor label. The BLT column is our reproduction.

	without anchor label			with anchor label		
task	ridge	bilinear ridge (absolute)	BLT	PADRE	bilinear ridge (difference)	BLT
bulk modulus	64.9	56.4	**48.69**	57.6	**41.3**	46.99
Debye temperature	0.381	0.376	**0.3176**	0.34	**0.304**	0.3636
shear modulus	0.538	0.542	**0.4205**	0.485	0.50	**0.4186**
thermal conductivity	0.917	1.10	**0.8777**	0.975	1.09	**0.8655**
thermal expansion	0.329	**0.268**	0.3254	0.343	**0.235**	0.2798

Several evaluation controls materially affect apparent rankings. Uniform thinning of pair-censoring training from 1464 to 700 points costs ADT 1.0 mV but the cubic 0.2 mV, because it removes local anchors; ladder folds sharing the same training set must be clustered (80 nominal pair-censoring folds are 40 training sets); local methods can look better by abstaining on hard points (the matched half of the ABC region under three-element isolation is 8.2% easier for every direct method); and pooled pointwise RMSE differs from mean per-fold RMSE (2.106 against 1.995 in σ units for the released MatEx code). We therefore use full training pools for within-scenario comparisons, report coverage for local methods, and bootstrap over independent training sets.