BEYOND FIT: INTERVENTION-AWARE HYPOTHESIS SEARCH FOR OOD TABULAR REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern tabular predictors perform strongly in-distribution but often degrade under out-of-distribution (OOD) shifts, especially when deployment requires extrapolation beyond observed support. A key difficulty is that fitting the source data does not uniquely determine how predictions should continue outside that support: different predictive rules can explain the observed data yet imply very different extrapolations. This suggests that OOD tabular regression requires not only learning or selecting a strong predictor, but also constructing alternative predictive hypotheses and deciding which of them should be used at deployment. We introduce *Intervention-Aware Hypothesis Search* (IAHS), a framework that separates these two problems by constructing executable hypotheses through representations, predictor families, predictive roles, and compositions, and then using source-derived evidence, permitted unlabeled deployment information, and deployment-specific constraints to guide protocol-conditioned adoption without target labels. On TabularMath, hypothesis construction increases the retrospective mean per-task best OOD R^2 from 0.3006 to 0.8177, while Source-score Top-1 selection on the same candidate pool reaches only 0.0625 compared with 0.5829 for IAHS. Across additional regression benchmarks, IAHS improves mean OOD R^2 over TabPFN v3 by 0.619 while maintaining comparable non-OOD performance. These results show that reliable extrapolation depends on both which predictive alternatives are made available and how they are selected under the evidence available at deployment.

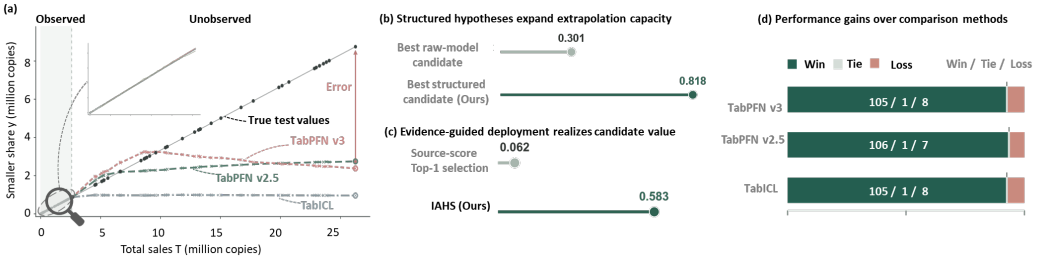


Figure 1: **Motivation for Intervention-Aware Hypothesis Search (IAHS). (a)** On TabularMath case `GSM8K-000088`, the evaluated models closely fit the source range (shaded region, enlarged in the inset), yet underestimate responses at larger unseen inputs. Dark points denote ground-truth test values, and the arrow marks TabPFN v3's prediction error. **(b)** Retrospective oracle capacity of raw models and constructed hypotheses, measured by mean per-case best R^2. **(c)** End-to-end OOD performance of source-score Top-1 selection and protocol-conditioned IAHS. **(d)** IAHS's pairwise R^2 win, tie, and loss counts against baselines. Panels **(b)–(d)** use TabularMath OOD cases.

1 INTRODUCTION

Tabular data underlie quantitative prediction across science and industry, from estimating material properties to forecasting economic outcomes Hollmann et al. (2025); Gorishniy et al. (2025). Recent tabular learners, including pretrained foundation models, have achieved strong performance when test data resemble the training distribution. In many practical settings, however, predictions are

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.

required under distribution shift and sometimes beyond the support of the observed data Cheng et al. (2026). Under such extrapolative shifts, strong in-distribution performance does not necessarily translate into reliable predictions outside the training range. Fig. 1(a) illustrates this gap: several strong tabular predictors closely fit the source observations yet incur substantial errors at larger unseen inputs.

A common response to distribution shift is to develop stronger predictors, larger pretrained models, or more effective model-selection and ensembling strategies. These approaches are effective when a suitable prediction rule is already represented among the available candidates. Extrapolation poses an additional difficulty: the relationship that remains useful outside the observed support may not be expressed by any predictor trained directly on the original representation. Our analysis in Fig. 1(b) illustrates this limitation: even retrospectively selecting the best raw-feature model leaves a substantial gap to predictions obtainable from a richer set of constructed alternatives. Thus, the challenge is not only to choose among existing predictors, but also to determine what predictive rules should be considered in the first place.

These observations motivate a different view of OOD regression. Outside the observed support, choices of input representation, predictor family, predictive role, and composition encode different assumptions about how relationships observed in the source data should extend into unseen regimes. We therefore view these choices as ways of constructing alternative *extrapolation hypotheses*, rather than merely adding features or models. Expanding the hypothesis space, however, creates a second difficulty: the target labels needed to identify the appropriate continuation are unavailable at deployment. Source-side evidence can therefore favor a candidate that fails under the actual shift, as illustrated by Source-score Top-1 selection in Fig. 1(c). OOD tabular regression consequently involves two distinct but coupled problems: constructing useful predictive alternatives and determining which alternative should be used given the evidence available at deployment.

We address these problems with *Intervention-Aware Hypothesis Search* (IAHS), a framework for constructing predictive alternatives and adapting their adoption to the information available under different deployment protocols. IAHS constructs executable predictive hypotheses by varying input representations, predictor families, predictive roles, and compositions. It then evaluates these hypotheses using source-derived evidence, permitted unlabeled deployment information, and task constraints specified by the deployment protocol. A protocol-conditioned adoption rule determines which complete hypothesis should produce the final prediction. Different deployment protocols can therefore instantiate the adoption stage differently while sharing the same hypothesis-construction and evidence-guided adoption interface. Thus, constructing a candidate does not by itself justify using it.

We evaluate IAHS on TabularMath and additional regression benchmarks spanning multiple application domains. On TabularMath, IAHS substantially improves OOD performance over strong tabular baselines. On the additional regression benchmarks, IAHS achieves a mean OOD R^2 of 0.2474, compared with -0.3716 for TabPFN v3, corresponding to an improvement of 0.619. At the same time, their mean non-OOD scores remain comparable, at 0.8870 and 0.8898, respectively. Further analyses distinguish the predictive opportunity created by hypothesis construction from the performance realized through evidence-guided adoption, and examine which prediction rules contribute to the final results.

Our contributions are threefold:

- **A hypothesis-search view of OOD tabular regression.** We formulate extrapolation as jointly constructing predictive alternatives and deciding which should be used at deployment, rather than only selecting among a fixed set of predictors.
- **Intervention-Aware Hypothesis Search.** We introduce IAHS, a framework that constructs executable hypotheses through representations, predictor families, predictive roles, and compositions, and supports protocol-conditioned evidence-guided adoption across deployment settings.
- **Evaluation across diverse OOD regression settings.** We evaluate IAHS on TabularMath and additional regression benchmarks spanning multiple application domains, separating the predictive opportunity created by hypothesis construction from the performance realized through evidence-guided adoption.

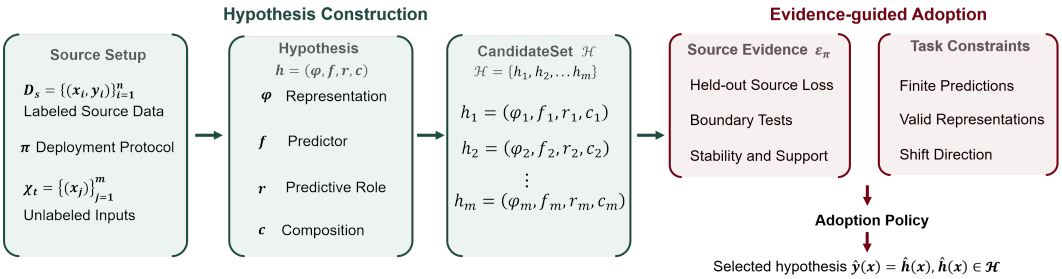


Figure 2: **Overview of Intervention-Aware Hypothesis Search (IAHS). Left: Hypothesis construction.** Given labeled source data $\mathcal{D}_s$, deployment protocol π, and optional permitted unlabeled inputs $\mathcal{X}_t$, IAHS constructs a finite set of complete predictive hypotheses $h = (\phi, f, r, c)$ by varying representation, predictor, predictive role, and composition. **Right: Evidence-guided adoption.** Source-derived environments $\mathcal{E}_\pi$ provide validation, boundary, stability, and support evidence, while task constraints encode validity and shift information. A protocol-conditioned adoption rule uses these quantities to select a complete hypothesis $\widehat{h}$, yielding $\widehat{y}(\boldsymbol{x}) = \widehat{h}(\boldsymbol{x})$. Target responses are not used for hypothesis construction or adoption.

2 Related Work

2.1 Tabular Learning

Recent tabular learning methods include strong tree-based learners, specialized deep architectures, retrieval-augmented predictors, and pretrained tabular foundation models. TabPFN performs in-context tabular prediction from synthetic pretraining tasks, with later variants improving scalability and regression performance (Hollmann et al., 2025; Grinsztajn et al., 2025; 2026). TabR combines deep tabular learning with nearest-neighbor retrieval, while TabM demonstrates the effectiveness of parameter-efficient ensembling (Gorishniy et al., 2024; 2025). These methods substantially improve standard tabular prediction, but strong held-out performance does not uniquely determine behavior under extrapolation. TabularMath makes this gap explicit by evaluating computational extrapolation in tabular regression (Cheng et al., 2026). We focus instead on how alternative predictive hypotheses should be constructed and selectively used when source fit alone is insufficient.

2.2 Structural Hypotheses and OOD Model Selection

Structural approaches such as symbolic regression introduce explicit inductive biases through compact functional relationships or transformed representations (Udrescu & Tegmark, 2020). Yet predictors with similar source-domain performance can behave substantially differently under distribution shift, as highlighted by work on underspecification (D'Amour et al., 2022). Related work in domain generalization and selective prediction further shows that model selection itself becomes difficult when deployment conditions differ from training (Gulrajani & Lopez-Paz, 2020; Geifman & El-Yaniv, 2017). IAHS differs by separating hypothesis construction from deployment authority: it constructs executable predictive alternatives and then asks whether available pre-deployment evidence is sufficient for one of them to modify a protected incumbent.

3 Intervention-Aware Hypothesis Search

A good fit on observed data does not uniquely determine how a predictor should continue beyond the observed range. IAHS addresses two linked problems: constructing useful extrapolation hypotheses and deciding which of them should be used at deployment. As illustrated in Fig. 2, these two problems lead to two stages: hypothesis construction and evidence-guided adoption. The first stage constructs a finite set of complete prediction rules by varying input representations, predictors, predictive roles, and compositions. The second stage evaluates these candidates using source-derived evidence, permitted unlabeled deployment information, and the constraints of the deployment setting, and selects one complete rule without using target responses. Thus, hypothesis construction

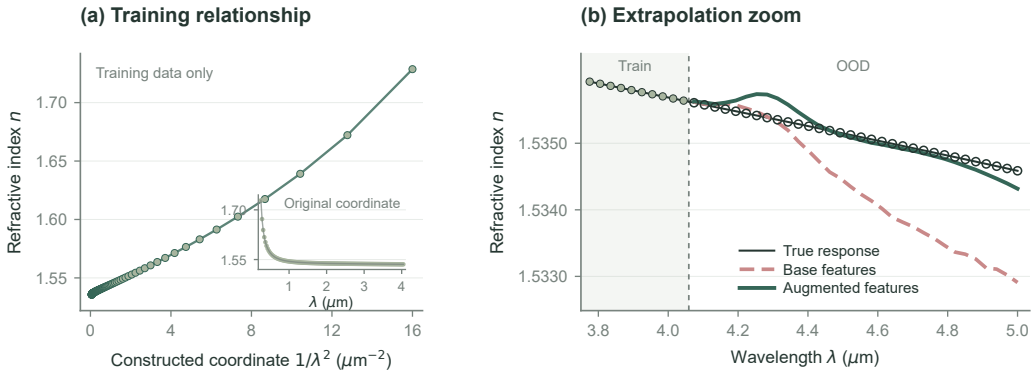


Figure 3: **Representation changes extrapolation: a KBr example. (a)** The same 128 training observations of potassium bromide (KBr) refractive index are shown against $1/\lambda^2$ and, in the inset, the original wavelength λ. **(b)** With the split and TabPFN v2.5 configuration fixed, augmenting the base features $(\lambda, \lambda^2, \log \lambda)$ with 20 source-derived terms changes the continuation beyond the training boundary and improves OOD R^2 from -9.12 to 0.74. Held-out responses are used only for evaluation; the comparison reflects the full augmented representation rather than $1/\lambda^2$ alone.

determines which predictive alternatives are available, while evidence-guided adoption determines which of them should produce the final prediction.

3.1 Problem Setup

For a regression task, we are given labeled training data $\mathcal{D}_s = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, which we refer to as the source data, together with unlabeled deployment inputs $\mathcal{X}_t = \{\boldsymbol{x}_j^t\}_{j=1}^{n_t}$ for which predictions are required. IAHS constructs predictive hypotheses using the source data and selects the final hypothesis using source-side validation evidence together with information from $\mathcal{X}_t$. The corresponding target responses remain unobserved during both hypothesis construction and selection and are used only for final evaluation.

We use π to denote the deployment protocol, which specifies the information available at selection time and any task-specific information provided by the evaluation setting. For example, the *target-high* protocol specifies extrapolation toward response values above the observed training range, whereas the *independent and identically distributed* (IID) setting provides no systematic directional shift. We denote the corresponding task constraints by $\mathcal{C}_\pi$, which include prediction validity, admissible representations, and compatibility with shift-direction information when such information is available. Together with the selection evidence, these constraints guide the choice of a complete prediction rule $\widehat{h}$, yielding $\widehat{y}(\boldsymbol{x}) = \widehat{h}(\boldsymbol{x})$.

3.2 Hypothesis Construction

With the available information defined, we next describe how IAHS constructs the predictive alternatives considered for adoption. A predictor's extrapolation behavior depends not only on the model itself but also on how the inputs are represented and how predictions are constructed. Restricting the candidate space to models fitted on a fixed representation can therefore leave useful continuations unexplored. IAHS addresses this limitation by constructing complete predictive hypotheses that vary the representation, predictor, predictive role, and composition. Fig. 3 illustrates this effect: changing the representation can substantially alter the predicted continuation beyond the training boundary even when the predictor configuration is held fixed.

We call each complete input-to-prediction rule a *predictive hypothesis*. Following Fig. 2, we write $h = (\phi, f, r, c)$, where ϕ denotes the input representation, f the fitted predictor, r the predictive role of constructed information, and c any prediction-level composition. A transformed feature or structural relation alone is therefore not a hypothesis; it becomes part of one only when incorporated into a complete prediction rule.

IAHS begins with predictors trained on the original features, including linear models, tree-based methods, and pretrained tabular models. It then constructs additional representations from the source data using a bounded set of transformations and interactions, including powers, logarithmic transformations, products, and safeguarded ratios, while discarding numerically invalid or unstable features. Given a constructed representation $\phi(\boldsymbol{x})$, IAHS can form a direct rule $h(\boldsymbol{x}) = f(\phi(\boldsymbol{x}))$ or augment the original inputs through $h(\boldsymbol{x}) = f([\boldsymbol{x}, \phi(\boldsymbol{x})])$. The framework also supports residual roles in which a second predictor corrects an initial prediction, although this option is not used in the primary target-high configuration. These choices allow the same predictor family to define different extrapolation hypotheses through different representations and predictive roles.

IAHS can further construct hypotheses by correcting or combining complete prediction paths. For example, source-derived bias or affine corrections can adjust prediction level or scale, while combinations of several paths can exploit complementary estimates. Task-specific information can also introduce additional candidates; in the target-high setting, for instance, the candidate set includes source-derived upper-tail rules that encode the known extrapolation direction. All such rules are constructed before adoption and evaluated as complete candidates. The resulting finite set is denoted

$$\mathcal{H}_\pi = \{h_1, \ldots, h_K\},$$

where hypotheses may differ in representation, predictor, predictive role, or composition. IAHS instantiates only the constructions supported by the current configuration rather than enumerating every possible combination. The next stage uses selection evidence and task constraints to determine which hypothesis should be adopted.

3.3 Selection Evidence and Task Constraints

After constructing $\mathcal{H}_\pi$, IAHS gathers the information used to decide which hypothesis should be adopted. As summarized in the right half of Fig. 2, this information is divided into *selection evidence* and *task constraints*.

Selection evidence. IAHS first evaluates complete candidate rules using internal splits of the labeled source data. Each candidate is fitted on one subset and evaluated on held-out source observations whose responses are known. In addition to ordinary validation, boundary and pseudo-OOD splits hold out selected source regions to probe whether a candidate can extend beyond the region on which it was fitted. IAHS also considers whether candidate performance is stable across these tests and whether the relevant inputs are sufficiently supported by the source data. All prediction errors used in these tests are computed from source responses.

The unlabeled deployment inputs provide an additional form of selection evidence. IAHS evaluates each candidate on $\mathcal{X}_t$ and examines the predictions it proposes on the deployment batch. Because the corresponding responses are unavailable, these predictions do not measure target accuracy; instead, they describe the continuation implied by the candidate. For example, in the target-high setting, IAHS examines whether the upper range of a candidate's predictions extends beyond the observed source response range. Large predicted values alone are not sufficient for adoption and are considered together with source validation, stability, and support evidence.

Task constraints. The task constraints $\mathcal{C}_\pi$ specify requirements that a candidate must satisfy before adoption. As summarized in Fig. 2, these include finite predictions, valid representations, and compatibility with a shift direction when such directional information is provided by the deployment protocol. For example, the target-high protocol supplies an upward-shift direction, whereas the IID protocol imposes no such directional requirement. The adoption policy described next combines the selection evidence with these constraints to choose the final predictive hypothesis.

3.4 Evidence-Guided Adoption

Having constructed the candidate set and gathered the corresponding selection evidence, IAHS applies an adoption policy to determine which complete hypothesis should produce the final predictions. The policy combines the available selection evidence with the task constraints rather than selecting a candidate by source-validation performance alone.

We use $\mathcal{E}_\pi$ to denote the selection evidence available under deployment protocol π. The final hypothesis is written as

$$\widehat{h} = \mathsf{Adopt}_\pi\left(\mathcal{H}_\pi, \mathcal{E}_\pi, \mathcal{C}_\pi\right), \qquad \widehat{y}(\boldsymbol{x}) = \widehat{h}(\boldsymbol{x}). \tag{1}$$

The selected hypothesis can be an original-input predictor, a rule using a constructed representation, or a previously evaluated correction or composition.

The specific adoption rule depends on the information available under the deployment protocol. When directional extrapolation information is available, candidate continuations can be checked for compatibility with that direction; when no such information is available, no directional requirement is imposed.

Here, an *intervention* refers to changing the prediction rule used for deployment rather than intervening on the underlying data-generating process.

4 Experiments

In this section, We first evaluate IAHS on TabularMath under both in-distribution and target-high OOD settings, and then on additional regression benchmarks covering diverse domains and distribution shifts. We further analyze the two components of IAHS by examining whether hypothesis construction produces stronger candidates and whether evidence-guided adoption can identify useful prediction rules without target labels.

4.1 Experimental Setup

TabularMath. We evaluate IAHS on 114 TabularMath problems, including 100 GSM8K-derived and 14 AIME-derived tables, following the official target-high evaluation protocol (Cheng et al., 2026). After applying a row cap of 2048, each problem is evaluated under two settings. The IID setting uses a random 80/20 training–test split as a control, while the official target-high OOD setting uses the lower-response 80% for training and the upper 20% for testing.

Additional regression benchmarks. We further evaluate IAHS on optical dispersion from RefractiveIndex.INFO (Polyanskiy, 2024), near-infrared regression using Tecator (Thodberg), molecular-property prediction using ESOL (Delaney, 2004) and FreeSolv (Mobley & Guthrie, 2014), and five TabArena datasets (Erickson et al., 2026). The reported scope contains 69 task–split cases, including 41 non-OOD cases and 28 OOD cases. Dataset coverage, split definitions, and scope details are provided in Appendices F and G.

Methods and metrics. We compare IAHS with Ridge regression (Hoerl & Kennard, 1970), tree-based methods ExtraTrees (Geurts et al., 2006), CatBoost (Prokhorenkova et al., 2018), and LightGBM (Ke et al., 2017), pretrained tabular models TabPFN v2.5 (Grinsztajn et al., 2025), TabPFN v3 (Grinsztajn et al., 2026), and TabICL v2 (Qu et al., 2026), and neural tabular models TabM (Gorishniy et al., 2025) and RealMLP (Holzmüller et al., 2024). TabM and RealMLP are external baselines and are not included in the IAHS candidate set.

All evaluations follow the information setting defined in Sec. 3.1. For target-high tasks, we additionally evaluate fixed predictors with a source-maximum floor to control for the upward-shift information available to IAHS. Implementation details and adoption settings are provided in Appendices B, C.1, and C.2.

We report case-level test R^2, its mean and median across cases, and paired win/tie/loss counts (W/T/L). All cases are weighted equally, with $|\Delta R^2| \leq 10^{-8}$ counted as a tie.

4.2 Main Results on TabularMath

Tab. 1 summarizes OOD and IID performance across all 114 TabularMath problems.

OOD Performance. IAHS achieves a mean OOD R^2 of 0.5829 and a median of 0.9981. Against the pretrained tabular baselines, it wins 106 of 114 cases against TabPFN v2.5 and 105 cases against both TabPFN v3 and TabICL v2. The improvement is therefore reflected not only in aggregate performance but also across individual tasks.

Table 1: **TabularMath IID and target-high OOD performance.** Each available entry contains all 114 cases of the corresponding split. Scores are test R^2, and W/T/L compares IAHS with the method in that row.

	IID		OOD		
Method	Mean	IAHS W/T/L	Mean	Median	IAHS W/T/L
Ridge	0.8426	99/8/7	-1.8859	-0.8926	92/15/7
ExtraTrees	0.9108	105/0/9	-5.1002	-2.5059	111/1/2
CatBoost	0.9369	105/0/9	-4.4510	-2.0103	110/1/3
LightGBM	0.9305	106/0/8	-3.4803	-1.6639	111/1/2
TabPFN v2.5	0.9680	50/8/56	-0.5071	0.2303	106/1/7
TabPFN v3	0.9672	73/6/35	-0.7378	0.1230	105/1/8
TabICL v2	0.9577	69/0/45	-1.2262	-0.0938	105/1/8
TabM	0.9573	102/0/12	-1.0330×10^{11}	0.3940	106/0/8
RealMLP	0.9287	107/0/7	-1.4595×10^{11}	-1.4705	110/0/4
IAHS	**0.9693**	–	**0.5829**	**0.9981**	–

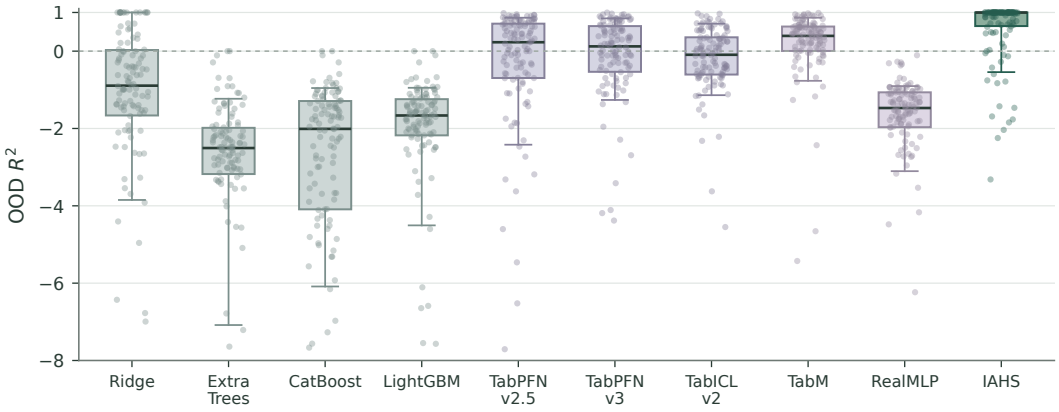


Figure 4: **OOD performance across TabularMath tasks.** Test R^2 distributions over the 114 target-high OOD tasks for IAHS and nine raw-input tabular baselines. Each visible point denotes one task; boxes show the interquartile range and median, and whiskers denote the 10th and 90th percentiles. For readability, the vertical axis is limited to $R^2 \geq -8$; all scores are retained when computing the reported statistics.

Fig. 4 shows the corresponding case-level distributions. Several baselines exhibit substantial negative tails, so their means can be strongly affected by a small number of severe failures even when their medians are considerably higher. IAHS shifts most of the distribution toward high R^2, while retaining a small number of difficult cases. We therefore report means, medians, and paired comparisons together.

IID performance. As shown in Tab. 1, IAHS achieves the highest mean IID R^2 among the compared methods, with 0.9693, compared with 0.9680 for TabPFN v2.5, 0.9672 for TabPFN v3, and 0.9577 for TabICL v2. This shows that IAHS improves OOD performance while preserving strong in-distribution accuracy.

4.3 Evaluation Beyond TabularMath

We next evaluate IAHS on the additional regression benchmarks introduced in Sec. 4.1. Table 2 summarizes performance on the 41 non-OOD and 28 OOD cases.

OOD performance. IAHS achieves a mean OOD R^2 of 0.2474 and a median of 0.6396. It wins 26 of 28 cases against TabPFN v3 and TabM, 24 against TabICL v2, and 21 against TabPFN v2.5. The gap between mean and median for several methods also reveals substantial variation in failure severity across tasks. Domain-level comparisons show particularly consistent improvements on op-

Table 2: **Performance beyond TabularMath.** All methods are evaluated on the same 41 non-OOD and 28 OOD cases. Scores are test R^2, and W/T/L compares IAHS with each baseline on the OOD cases.

Method	Non-OOD mean	OOD mean	OOD median	IAHS W/T/L
Ridge	0.6856	-6864.7971	-0.0685	25/3/0
ExtraTrees	0.8308	-21.2441	-3.1526	27/0/1
CatBoost	0.8148	-23.2056	-2.5694	27/0/1
LightGBM	0.7934	-14.2669	-2.5529	27/0/1
TabPFN v2.5	0.8806	-4.1759	-0.1724	21/0/7
TabPFN v3	**0.8898**	-0.3716	-0.1533	26/0/2
TabICL v2	0.8876	-511.7361	-0.3242	24/0/4
TabM	0.8393	-744.9533	-0.5971	26/0/2
RealMLP	0.8425	-11.2110	-1.8266	27/0/1
IAHS	0.8870	**0.2474**	**0.6396**	–

Table 3: **Selected ablations on TabularMath.** All settings use the same 114 IID and 114 OOD cases. Source-score Top-1 changes selection on the same candidate pool. The remaining variants independently restrict available prediction rules and repeat selection without refitting the underlying predictors. Full − setting and W/T/L compare full IAHS with each variant on OOD. Complete ablations are reported in Appendix D.

Setting	IID mean	OOD mean	Full − setting	Full W/T/L
Full IAHS	0.9693	0.5829	–	–
Source-score Top-1	0.9699	0.0625	+0.5204	71/16/27
Direct rules only	0.9700	0.0519	+0.5309	79/35/0
w/o source-quantile floor	0.9693	0.4565	+0.1264	70/44/0
w/o bias/affine or descendants	0.9693	0.1784	+0.4045	9/105/0

tical dispersion, where IAHS outperforms TabPFN v3 on all 20 OOD cases. Complete results by application domain are provided in Appendix G.

Non-OOD performance. On the 41 non-OOD cases, IAHS obtains a mean R^2 of 0.8870, compared with 0.8898 for TabPFN v3, 0.8876 for TabICL v2, and 0.8806 for TabPFN v2.5. Thus, the additional benchmarks show substantially different behavior under OOD evaluation while IAHS remains closely matched with the strongest baselines on the non-OOD cases. Together, these results indicate that the benefits of hypothesis construction and adoption extend beyond TabularMath, while remaining dependent on the task and shift setting.

4.4 Analysis and Ablations

We analyze the two core stages of IAHS on the same 114 IID and 114 OOD TabularMath cases as the main evaluation. We first ask whether hypothesis construction provides stronger predictive alternatives and whether evidence-guided adoption can make effective use of them. We then examine which prediction rules contribute to the final results.

Does construction provide stronger candidates? Fig. 1(b) compares candidate capacity retrospectively. For each case, we take the highest test R^2 within each candidate pool and average these per-case best scores. The mean increases from 0.3006 for the seven-model raw-input pool to 0.8177 for the constructed hypothesis pool. Thus, hypothesis construction substantially expands the set of useful predictive alternatives. This is a reporting-only oracle comparison that measures candidate-pool capacity rather than deployable selection performance.

Does the adoption policy matter? We next keep the candidate pool fixed and compare IAHS with *Source-score Top-1*, which selects the candidate with the highest recorded source-side score. Neither policy uses target responses for selection. Source-score Top-1 achieves a mean OOD R^2 of 0.0625, compared with 0.5829 for IAHS. Across the same 114 OOD cases, IAHS records 71/16/27 wins/ties/losses. Thus, making stronger candidates available is not sufficient. How those candidates are selected at deployment is also critical to OOD performance.

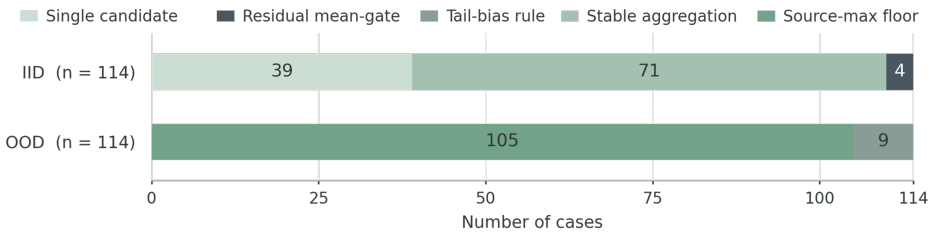


Figure 5: **Selected rule families under IID and target-high OOD evaluation.** Each bar contains all 114 TabularMath cases of the corresponding setting. Segment counts indicate the final selected rule family. Selection frequencies characterize adoption behavior rather than the performance contribution of individual rule families.

Which prediction rules matter? We restrict the available rules while keeping the fitted predictors and constructed representations fixed. Direct rules use a single prediction path without subsequent output correction or aggregation. Floor rules lower-bound predictions by source-response quantiles, while bias and affine rules shift or rescale predictions using source-side residuals. Detailed definitions are provided in Appendix D.

Using direct rules only reduces the mean OOD R^2 from 0.5829 to 0.0519, showing that constructed representations alone do not explain the full gain. Removing source-quantile floors reduces the mean to 0.4565, while removing bias/affine rules and their descendants reduces it to 0.1784. These results show that prediction-level corrections make an important contribution to IAHS.

Because source-maximum floors are frequently adopted in target-high tasks, we also compare against a fixed degree-two sparse polynomial predictor equipped with the same source-maximum floor. This strong control reaches a mean OOD R^2 of 0.5642 and a median of 0.8910, compared with 0.5829 and 0.9981 for IAHS. IAHS also records 69/29/16 wins/ties/losses against this control. Thus, the target-high direction provides useful information, while IAHS achieves stronger aggregate performance through a broader set of constructed and corrected prediction rules. Complete controls are reported in Appendix D.5.

How does adoption differ across settings? Fig. 5 shows that IAHS uses different parts of the hypothesis space under IID and target-high OOD evaluation. IID cases mainly adopt stable aggregations and single-candidate rules, whereas target-high OOD cases primarily adopt source-maximum-floor and tail-bias rules. This difference provides direct evidence for setting-dependent adoption. The frequencies describe which rules are selected, while their performance effects are evaluated by the ablations above.

Overall, these analyses separate two complementary roles of IAHS. Hypothesis construction makes stronger predictive alternatives available, while evidence-guided adoption determines which alternatives are used under each deployment setting.

5 Conclusion

Modern tabular predictors can perform strongly in-distribution, yet extrapolative OOD prediction remains challenging because source fit alone does not determine how predictions should continue beyond the observed support. We introduced *Intervention-Aware Hypothesis Search* (IAHS), a framework that separates the construction of alternative predictive hypotheses from their evidence-guided adoption at deployment. IAHS explores hypotheses across representations, predictor families, predictive roles, and compositions, and selects among them without access to target responses. Across TabularMath and additional regression benchmarks, IAHS improves OOD performance over strong tabular baselines, including a 0.619 gain in mean OOD R^2 over TabPFN v3 on the additional benchmarks. These results highlight both hypothesis construction and evidence-guided adoption as important for reliable tabular extrapolation.

AI USE STATEMENT

Generative AI tools were used selectively during this work to support code implementation and manuscript preparation. All AI-assisted outputs were inspected, revised, and validated by the authors before use. The reported experiments and scientific conclusions are based on the authors' implementation and analysis. The authors take full responsibility for the final manuscript.

ETHICS STATEMENT

This work studies machine-learning methods on existing tabular regression benchmarks and does not involve human subjects, personally identifiable information, or interventions on individuals. We are not aware of direct societal risks specific to the proposed method beyond those associated with deploying regression models under distribution shift. To support transparent evaluation, benchmark protocols, reporting groups, and post-hoc scope revisions are explicitly documented in the appendix.

REPRODUCIBILITY STATEMENT

We provide detailed information for reproducing the reported experiments in the main paper and appendix, including the IAHS construction and adoption procedures, experimental configurations, dataset and split protocols, baseline settings, ablations, evaluation metrics, and random seeds. Code and experiment configurations for reproducing the main results will be released publicly.

REFERENCES

Zerui Cheng, Jiashuo Liu, Jianzhu Yao, Pramod Viswanath, Ge Zhang, and Wenhao Huang. Tabular-math: Evaluating computational extrapolation in tabular learning via program-verified synthesis. In *3rd AI for Math Workshop: Toward Self-Evolving Scientific Agents at ICML 2026*, 2026.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems*, 38, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. Tabr: Tabular deep learning meets nearest neighbors. In *International Conference on Learning Representations*, volume 2024, pp. 18209–18249, 2024.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. Tabm: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, volume 2025, pp. 77899–77935, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned mlps and boosted trees on tabular data. In *Neural Information Processing Systems*, 2024.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28(7):711–720, 2014.

Mikhail N Polyanskiy. Refractiveindex. info database of optical constants. *Scientific Data*, 11(1): 94, 2024.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

Jingang Qu, David HolzmÃžller, GaÃl Varoquaux, and Marine Le Morvan. Tabiclv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Hans Henrik Thodberg. Tecator data set. StatLib Datasets Archive, Carnegie Mellon University. URL `https://lib.stat.cmu.edu/datasets/tecator`.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science advances*, 6(16):eaay2631, 2020.

A Additional Visualizations

This section provides visual complements to the main results, including TabularMath IID performance and domain- and case-level views of the additional OOD benchmarks. All figures use the same predictions and evaluation scopes as the corresponding result tables.

A.1 TabularMath in-distribution performance

Figure 6 shows all 114 TabularMath tasks under the IID split. It complements Table 1 by showing the case-level performance distribution.

B Implementation Details

B.1 Primary TabularMath Configuration

The primary TabularMath evaluation contains 114 IID and 114 target-high OOD cases. Each table is capped at 2048 rows before splitting. We use split-generation seed 2025 and model seed 42. All hypothesis construction and source-side evidence use only the official training partition. Unlabeled deployment inputs may be used for support and prediction summaries, while target responses are never available to the executed adoption policy.

Source-side candidate evaluation uses a boundary holdout with validation fraction 0.25 and at least 24 validation examples, together with up to three additional source-derived validation environments when available.

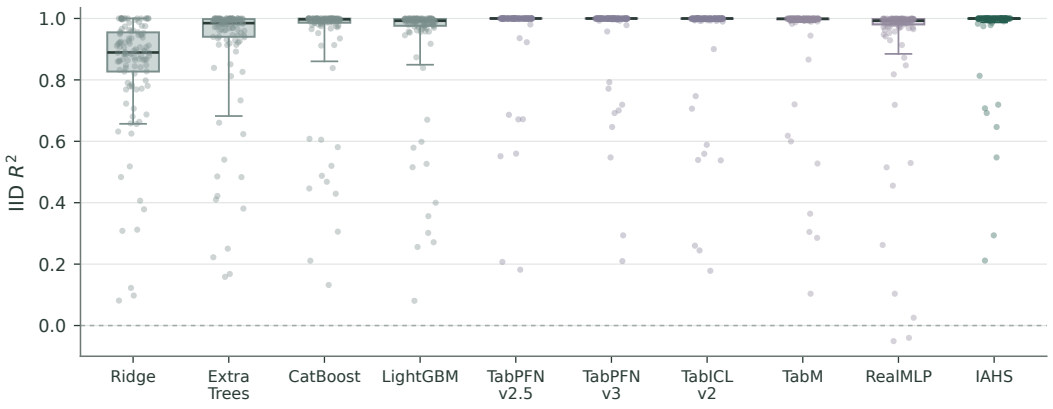


Figure 6: **IID performance across all 114 TabularMath tasks.** IAHS and nine raw-input baselines are evaluated on the same tasks. Each point represents one task. Boxes show the interquartile range and median, and whiskers show the 10th and 90th percentiles. All scores are shown on the same unclipped linear scale.

Table 4: **Configuration of the primary TabularMath experiments.**

Setting	Value
LLM mode	Off
Row cap	2048
Split / model seed	2025 / 42
Primary inner validation	Boundary holdout
Validation fraction / minimum size	0.25 / 24
Additional source environments	At most 3
Additional validation fraction	0.22
Cross-environment standard-deviation penalty	0.35
TabPFN v2.5 / v3 / TabICL v2 estimators	4 each
ExtraTrees estimators / minimum leaf size	512 / 2
Constructed-feature budget	14
Sparse-regression CV folds	5
Sparse-regression alpha count / iteration cap	24 / 30,000
Diagnostic-triggered degree-three regression	Permitted
Symbolic base-variable / pair-variable budgets	8 / 5
Source-response floor quantiles	0.90, 0.95, 1.00
Affine slope range	[0.25, 3.00]
Tail prediction quantile / margin	0.90 / 0.05
Target-high residual correction	Disabled

The effective source environments depend on the deployment protocol and construction plan. Some representation proposals use the full source partition before inner validation; their resulting scores are therefore used as selection evidence rather than interpreted as fully nested estimates.

B.2 Candidate Construction Details

IAHS constructs additional representations from a bounded collection of transformations and interactions, including powers, logarithmic transformations, products, and safeguarded ratios. Numerically invalid or unstable features are removed before fitting.

Constructed representations may be used directly or appended to the original inputs. The resulting prediction paths can further be used to construct source-derived bias and affine corrections, response floors, and prediction-level compositions. The target-high configuration additionally permits upper-tail rules motivated by the directional information supplied by the protocol. All resulting candidates are complete prediction rules before adoption.

B.3 Baseline Configurations

The supplementary CatBoost and LightGBM evaluations use fixed hyperparameters across all TabularMath tasks. Both use 500 boosting iterations, learning rate 0.035, model seed 42, and four CPU threads. CatBoost uses depth 6, random strength 0.25, and ℓ_2 leaf regularization 5. LightGBM uses 31 leaves and ℓ_2 regularization 2. No task-specific hyperparameter tuning is performed.

TabM and RealMLP are evaluated as external baselines and are not included in the IAHS candidate pool.

C Adoption Policies

C.1 Primary TabularMath Adoption

The primary TabularMath experiments instantiate the general IAHS adoption interface with separate policies for target-high OOD and IID evaluation. Both policies use source-derived evidence and candidate predictions on the unlabeled deployment inputs. Target responses are not used for adoption.

Source scores and deployment summaries. Let $s(h)$ denote the recorded source-side score of candidate rule h. For target-high evaluation, define

$$m_s = \max_i y_i, \qquad d_s = \max\left\{10^{-9}, m_s - Q_{0.9}(\{y_i\}_{i=1}^n)\right\}. \tag{2}$$

The upper-tail prediction summary of candidate h on the unlabeled deployment inputs is

$$q_t(h) = Q_{0.9}\left(\{h(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}_t\}\right), \tag{3}$$

and its normalized extrapolation level is

$$\ell_t(h) = \frac{q_t(h) - m_s}{d_s}. \tag{4}$$

These quantities use source responses and candidate predictions on unlabeled deployment inputs, not target responses. The standard tail-capability check is

$$q_t(h) \geq m_s - 0.05 d_s. \tag{5}$$

Corrections and compositions. For an underlying prediction path, a bias rule adds the median residual estimated on its inner source-validation partition. An affine rule fits a source-side slope and intercept, with the slope clipped to $[0.25, 3]$. Floor rules use source-response quantiles $q \in \{0.90, 0.95, 1.00\}$, computed from the fitting partition during validation and from the full source training partition at deployment. Configured quantile compositions can further combine source-competitive bias-corrected prediction paths. All corrections and compositions are constructed before final adoption.

Target-high adoption. For valid prediction paths, the source-competitive router first retains those whose source-validation MAE is at most 1.5 times the best available MAE. Within this set, it favors candidates with stronger upper-tail behavior on the unlabeled deployment batch. Admission to the preferred target-high branch additionally requires sufficient source-side evidence and compatibility with the target-high direction.

The final target-high policy gives priority to eligible routed rules and their supported source-maximum-floor or correction variants. If these are unavailable, it considers other tail-capable floor and composition rules before falling back to the remaining source-supported candidates. A bounded symbolic-tail branch can replace an extreme routed continuation when a strongly source-supported sparse-polynomial correction provides a more moderate compatible continuation. All priority rules and thresholds are fixed before target evaluation.

For eligible source-maximum floors, the policy compares their deployment prediction geometry relative to the source response range. If no preferred branch applies, the structural fallback orders routed rules, source-maximum floors, bias corrections, and affine corrections before reverting to the remaining source-scored upper-tail rules.

IID adoption. The target-high branch is not invoked for IID tasks. IID adoption instead emphasizes source-side performance and stability without imposing a directional requirement on deployment predictions.

The policy begins from the strongest eligible source-supported rule. For more difficult source-side cases, pseudo-OOD-supported candidates and stable aggregations can replace the initial choice when their scores remain sufficiently competitive. Residual mean-gate rules require an additional source-side improvement before adoption. These preferences are fixed independently of target responses and allow IID and target-high evaluations to adopt different parts of the same constructed hypothesis space.

C.2 Reference-Based Adoption for Additional Benchmarks

The additional regression benchmarks use the same hypothesis-construction and evidence-guided adoption interface, but do not use the target-high hierarchy above. Instead, the adoption stage evaluates complete alternatives relative to a frozen reference using source-side evidence matched to the deployment covariates.

Target-matched source environments. A cross-fitted logistic domain classifier distinguishes source inputs from the unlabeled deployment inputs. Its binary labels indicate domain membership rather than regression responses. Estimated propensities rank source observations by similarity to the deployment covariates.

When the available data and diagnostics support doing so, high-propensity source subsets form additional validation environments, with the remaining source observations used for fitting. Every regression loss used by the adoption rule is still evaluated using source responses. Additional environments are omitted when the available data do not support their construction.

Reference and challenger evidence. The policy first selects and freezes a reference hypothesis b from the core complete proposals. The reference need not be a raw-feature predictor and may itself contain a composition. For challenger h, its relative gain on source environment e is

$$g_e(h, b) = L_e(b) - L_e(h). \tag{6}$$

A positive value indicates improvement over the reference on that source-side condition and does not represent observed target improvement.

Adoption. The policy combines target-matched and aggregate source-environment losses with cross-environment stability, predictor disagreement, prediction validity, shift compatibility, and complexity information. A challenger is adopted only when the available source-side evidence supports changing the frozen reference. Otherwise, the reference prediction is retained.

A selected proposal that overlaps the reference prediction paths can produce an evaluated combination, while a distinct proposal can replace the reference. Both are complete hypotheses constructed before final adoption.

This reference-based policy is one realization of the general IAHS interface. It is distinct from the target-high hierarchy used in the primary TabularMath experiment.

C.3 Instantiation Across Evaluation Settings

The primary TabularMath experiment and the additional benchmarks share the same complete-hypothesis construction and evidence-guided adoption abstraction but instantiate the adoption stage differently according to the available deployment information.

Thus, the two evaluations do not apply one fixed selector. They are protocol-specific realizations of the same IAHS abstraction: hypothesis construction determines which complete prediction rules are available, while evidence-guided adoption determines which of them is used under the information available at deployment.

Table 5: **IAHS instantiations across evaluation settings.** Both settings construct and select complete prediction rules without using target responses at deployment.

Aspect	Primary TabularMath	Additional benchmarks
Task information	Target-high response direction for OOD; no directional constraint for IID	Task-specific protocol information such as wavelength direction, scaffold hold-out, or designated spectroscopy splits
Deployment inputs	Input-support checks and candidate prediction summaries	Input-support checks and target-matched source environments when supported
Adoption	Protocol-conditioned target-high policy; source-score/stability policy for IID	Reference-based comparison of complete alternatives using matched source-side evidence
Output	Selected complete rule, including an evaluated correction or composition	Reference, evaluated combination, or alternative complete rule

D Supplementary Ablations and Controls

D.1 Complete Primary Ablations

Table 6 reports all completed current-configuration policy and rule-pool ablations. Each condition uses the same 114 IID and 114 target-high OOD cases. The fitted predictors and their representations are not retrained.

Table 6: **Complete TabularMath ablations.** W/T/L compares full IAHS with each ablation on the 114 OOD cases.

Setting	IID mean	OOD mean	OOD median	Full W/T/L
Full IAHS	0.9693	0.5829	0.9981	–
Source-score Top-1	0.9699	0.0625	0.9926	71/16/27
No protocol priority	0.9693	0.2920	0.6615	98/1/15
No symbolic-tail override	0.9693	0.1784	0.9977	9/105/0
Direct rules only	0.9700	0.0519	0.9971	79/35/0
w/o source-quantile floor	0.9693	0.4565	0.9977	70/44/0
w/o affine	0.9693	0.5829	0.9981	0/114/0
w/o bias or descendants	0.9693	0.5808	0.9977	7/105/2
w/o bias/affine or descendants	0.9693	0.1784	0.9977	9/105/0
w/o tail-quantile ensemble	0.9693	0.5829	0.9981	0/114/0
w/o multi-predictor aggregation	0.9700	0.5829	0.9981	0/114/0
w/o support wrapper	0.9693	0.5829	0.9981	0/114/0

D.2 Rule-Pool Intervention Definitions

Selection-policy interventions. Source-score Top-1 keeps the complete candidate pool and selects the candidate with the highest recorded source-side score. No protocol priority removes the target-high-specific final priority logic without changing the candidate pool. No symbolic-tail override disables the corresponding adoption branch while retaining the symbolic candidates and their other eligible rules.

Direct prediction paths. A prediction path consists of a fitted predictor and its representation, which may contain either raw or constructed features. The direct-rules-only condition retains these individual paths while removing subsequent calibration, floors, multi-predictor aggregation, and residual-correction rules. The final prediction path is reselected after this restriction rather than fixed to the path selected by full IAHS.

For target-high cases, the restricted router retains paths whose source-validation MAE is at most 1.5 times the best available MAE and selects according to their upper-tail predictions on the unlabeled deployment batch. For IID cases, the retained policy uses source-validation and pseudo-OOD evidence. Target responses enter neither selection procedure.

Source-quantile floors. For base prediction $p(\boldsymbol{x})$, a floor rule has the form

$$p_{\text{floor}}(\boldsymbol{x}) = \max\{p(\boldsymbol{x}), \tau_s\}, \tag{7}$$

where τ_s is a source-response quantile, including the source maximum. During validation the threshold is computed from the fitting partition, while deployment uses the full source training partition. Removing floors leaves independently constructed bias, affine, and aggregation rules available. A floor does not alter predictions already above its threshold.

Bias and affine corrections. Bias correction uses

$$p_{\text{bias}}(\boldsymbol{x}) = p(\boldsymbol{x}) + \delta, \tag{8}$$

where δ is estimated from the median source-validation residual. Affine correction uses

$$p_{\text{affine}}(\boldsymbol{x}) = a\, p(\boldsymbol{x}) + b, \tag{9}$$

with bounded slope and a source-derived intercept.

The joint bias/affine removal also removes aggregates constructed from bias-corrected paths. We refer to these derived rules as their descendants. Floor rules are constructed independently and therefore remain available.

Aggregation and support rules. Multi-predictor aggregation includes stable medians, pseudo-OOD stable medians, and configured quantile compositions. No tail-quantile ensemble removes only the corresponding bias-corrected quantile compositions. No multi-predictor aggregation removes all configured multi-path aggregates while retaining individual paths and their independently constructed corrections. No support wrapper removes only the support-based selection wrapper.

Each restriction is applied independently to the full rule pool and selection is rerun over the surviving rules. The effects of different removals should therefore not be interpreted as additive component contributions.

D.3 Selected-Rule Grouping

The 114 target-high OOD selections comprise 105 source-maximum-floor rules and nine tail-bias rules. The 114 IID selections comprise 69 stable medians, 36 pseudo-OOD-best rules, three validation-best rules, two pseudo-stable medians, and four residual mean-gate rules.

For Fig. 5, the IID selections are grouped into 71 stable aggregations, 39 single-candidate rules, and four residual mean-gate rules. These counts describe selected rule identities rather than the number of cases whose numerical predictions are changed by the corresponding rule.

D.4 Paired Uncertainty

For each comparison, we compute the 114 paired OOD case-level differences between IAHS and the comparator. We resample these paired cases with replacement 10,000 times using seed 20260924 and report the 2.5th and 97.5th percentiles of the resulting mean differences. No model is refitted.

Table 7: **Paired uncertainty on TabularMath OOD.** Values report IAHS minus the comparator.

Comparator	Mean gain	95% lower	95% upper
Source-score Top-1	0.5204	0.1847	0.9832
No source-quantile floor	0.1264	0.0225	0.2825
Sparse D2 + source-max floor	0.0187	-0.1196	0.1423

D.5 Protocol-Matched Fixed-Predictor Controls

To isolate a simple use of the target-high directional information, we equip fixed predictors with a source-maximum floor:

$$h^{\text{floor}}(\boldsymbol{x}) = \max\left\{h(\boldsymbol{x}), \max_{(\boldsymbol{x}_i, y_i)\in\mathcal{D}_s} y_i\right\}. \tag{10}$$

Table 8: **Fixed-predictor controls for target-high OOD.** All methods are evaluated on the same 114 OOD cases. W/T/L compares full IAHS with the floor-corrected predictor.

Fixed model	Raw mean	Floor mean	Floor median	IAHS W/T/L
Ridge	-1.8859	-0.5664	-0.5925	86/21/7
ExtraTrees	-5.1002	-1.6949	-1.5512	103/7/4
Sparse D1	-2.2293	-0.6581	-0.7641	100/8/6
Sparse D2	0.4133	0.5642	0.8910	69/29/16
TabPFN v2.5	-0.5071	-0.2437	0.2303	86/20/8
TabPFN v3	-0.7378	-0.0983	0.1446	98/6/10
TabICL v2	-1.2262	-0.2111	-0.0928	99/5/10

Table 9: **Additional datasets and evaluation protocols.** These are the tasks used in the reported evaluation rather than the full MoleculeNet or TabArena benchmarks.

Domain	Dataset or prediction target	Evaluation protocols
Molecular properties	Delaney; FreeSolv	Random 80/20; molecular-scaffold holdout; upper-response 20% holdout.
Near-infrared spectroscopy	Tecator fat and moisture	$C + M \to T$ interpolation; $C \to M$ designated holdout; $C + M \to E_1$ and $C + M \to E_2$ extrapolation.
Optical dispersion	Refractive index for ten material curves	Random 80/20; hold out the shortest or longest 20% of wavelengths.
TabArena subset	Airfoil self-noise, body density, concrete strength, Miami housing, wine quality	Five supplied repeat/fold pairs per dataset; no additional OOD split.

The floor level is determined entirely from source training responses and is not selected using target performance. Sparse D1 and Sparse D2 denote degree-one and degree-two sparse polynomial regression, respectively.

Sparse D2 with the source-maximum floor is the strongest of these fixed controls by mean OOD R^2. Its mean increases from 0.4133 to 0.5642 after applying the floor. IAHS reaches 0.5829, with a median of 0.9981 versus 0.8910 and 69/29/16 wins/ties/losses against this control.

The paired mean difference is 0.0187 and the interval in Table 7 includes zero. The mean, median, and paired counts therefore provide complementary views of this comparison.

E KBr Representation Diagnostic

The diagnostic in Fig. 3 uses 160 KBr refractive-index observations over wavelengths from 0.25 to 5.0 micrometers. The first 128 observations form the source set and the remaining 32 high-wavelength observations are held out. The training boundary is approximately 4.059 micrometers.

The baseline representation is $(\lambda, \lambda^2, \log \lambda)$. The augmented representation adds 20 source-derived terms, including $1/\lambda^2$. TabPFN v2.5, seed 42, four estimators, and the train–test split are held fixed. The held-out R^2 changes from -9.1186 to 0.7384.

The left panel of Fig. 3 shows the same source observations under two coordinate systems. The right panel displays the boundary region together with all held-out observations. Held-out responses are used only for retrospective evaluation and are not used for feature construction. The comparison evaluates the complete augmented representation and does not attribute the improvement to $1/\lambda^2$ alone. This diagnostic is separate from the additional benchmark suite and is not counted as an additional benchmark case.

F Additional Benchmark Protocols and Scope

The additional benchmarks evaluate separate regression tasks across multiple application domains. They do not train a model on TabularMath and transfer it to these datasets. The reported suite contains 69 task–split cases, divided into 41 non-OOD and 28 explicit OOD evaluations.

F.1 Dataset-Specific Protocols

Molecular properties. For Delaney and FreeSolv, the random split provides the non-OOD evaluation. Two OOD protocols are used for each dataset. The molecular-scaffold split groups molecules according to their Bemis–Murcko scaffolds, while the target-high split holds out the largest responses. Because scaffold groups are kept intact, the realized test fraction of the scaffold split need not equal exactly 20%. Together, the two datasets contribute two non-OOD and four OOD cases.

Near-infrared spectroscopy. The retained Tecator evaluation uses fat and moisture as prediction targets. The recorded subsets contain 129 calibration observations (C), 43 designated holdout observations (M), 43 interpolation-test observations (T), and two designated extrapolation subsets containing 8 (E_1) and 17 (E_2) observations.

Interpolation fits on $C + M$ and evaluates on T. The designated holdout evaluation fits on C and evaluates on M. The two extrapolation evaluations fit on $C + M$ and evaluate on E_1 and E_2, respectively. All four comparisons are retained for both fat and moisture.

The extrapolation subsets are fixed sample subsets and are reused across the two retained prediction targets rather than redefined according to each target response. These evaluations therefore reuse observations and should not be interpreted as independent datasets. The protein prediction target is excluded from both reporting groups as part of the scope revision described in Table **??**.

Optical dispersion. The optical suite contains ten refractive-index curves: Al_2O_3, AlN, BaB_2O_4, BaF_2, CaF_2, CsBr, CsCl, CsF, CsI, and KBr. Each curve contributes one random 80/20 interpolation split and two OOD splits obtained by holding out either the shortest or longest 20% of wavelengths. This gives 10 non-OOD and 20 OOD cases.

Wavelength is an input coordinate, so high- or low-wavelength extrapolation should not be interpreted as a response-direction split. The windowed KBr diagnostic in Appendix E is separate from these canonical benchmark cases.

TabArena. The TabArena subset contains airfoil self-noise, body-density prediction, concrete compressive strength, Miami housing, and wine quality. For each dataset, we use the supplied repeat/fold pairs $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, 0)$, and $(1, 1)$. These yield 25 non-OOD cases, not 25 independent datasets. No additional OOD split is introduced.

F.2 Non-OOD and OOD Reporting Groups

Table 10 summarizes the resulting evaluation scope. Case counts can include folds, targets, or shift directions within the same underlying dataset and therefore should not be interpreted as independent experimental replicates.

Table 10: **Scope of the additional regression benchmarks.** IDs denote dataset/target identifiers rather than independent task–split cases.

Domain	Non-OOD cases	OOD cases	Non-OOD IDs	OOD IDs
Molecular/QSAR	2	4	2	2
Near-infrared spectroscopy	4	4	2	2
Optical dispersion	10	20	10	10
TabArena regression subset	25	0	5	0
Total	41	28	19	14

The 41 non-OOD cases comprise 39 IID/interpolation evaluations and two designated Tecator holdout evaluations. Specifically, the 39 IID/interpolation cases contain ten optical, two molecular, two Tecator interpolation, and 25 TabArena evaluations, while the remaining two cases use the Tecator $C \to M$ holdout for fat and moisture prediction.

The 28 OOD cases comprise 20 optical wavelength-extrapolation cases, four molecular OOD cases, and four Tecator extrapolation cases. This grouping distinguishes explicit extrapolation protocols from the remaining evaluation settings without assuming that all non-OOD splits are identically distributed.

G Additional Benchmark Results

G.1 Results by Domain

Table 11: **Additional-benchmark results by domain and split.** Values are case-macro mean test R^2. TabPFN v3 is shown as a fixed pretrained-model reference.

Domain	Group	Cases	IDs	IAHS	TabPFN v3
Molecular/QSAR	Non-OOD	2	2	0.9334	0.9432
Molecular/QSAR	OOD	4	2	-0.5020	-0.9248
Near-infrared	Non-OOD	4	2	0.9962	0.9939
Near-infrared	OOD	4	2	-0.2526	-0.9695
Optical	Non-OOD	10	10	0.9977	0.9976
Optical	OOD	20	10	0.4972	-0.1414
TabArena subset	Non-OOD	25	5	0.8215	0.8258

The suite is not balanced by domain. Optical tasks account for 20 of the 28 OOD cases, while TabArena accounts for 25 of the 41 non-OOD cases. The aggregate means reported in the main text are therefore case-weighted summaries rather than equal-domain averages.

G.2 OOD Comparisons by Domain

Figure 7 separates the aggregate OOD comparison by application domain.

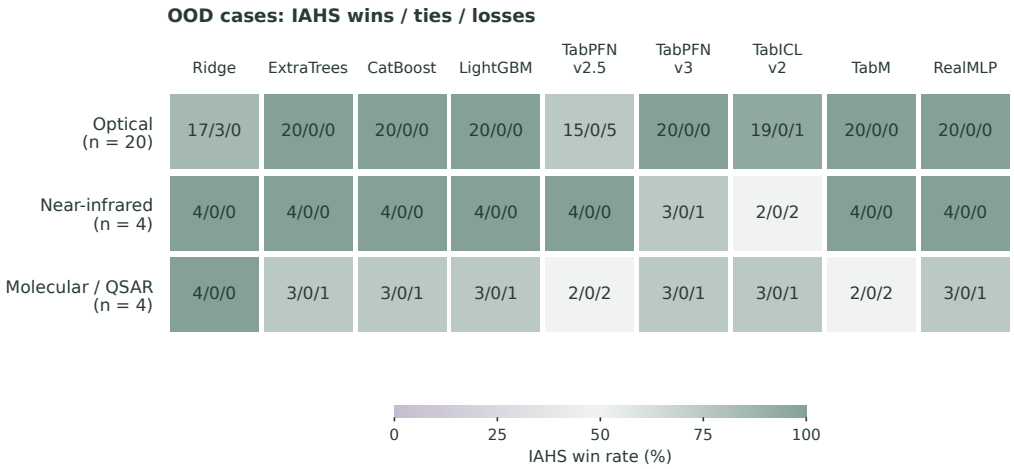


Figure 7: **OOD comparisons across application domains.** Each cell gives IAHS wins/ties/losses against a fixed baseline; ties satisfy $|\Delta R^2| \leq 10^{-8}$. Color encodes the within-domain win fraction rather than gain magnitude or statistical significance. All 28 OOD cases are included: 20 optical, four near-infrared, and four molecular/QSAR cases.

G.3 Case-Level Optical Extrapolation

The optical benchmark contributes the largest number of OOD cases. Figure 8 therefore shows every optical case in both extrapolation directions. TabPFN v3 is a fixed comparator rather than a baseline chosen separately for each case. IAHS improves on TabPFN v3 in all 20 paired optical comparisons.

G.4 Seed Sensitivity

The OOD median remains similar across the three recorded seeds, while the mean varies more because it is sensitive to a small number of difficult cases. These runs are sensitivity checks for the additional-benchmark realization rather than independent benchmark datasets.

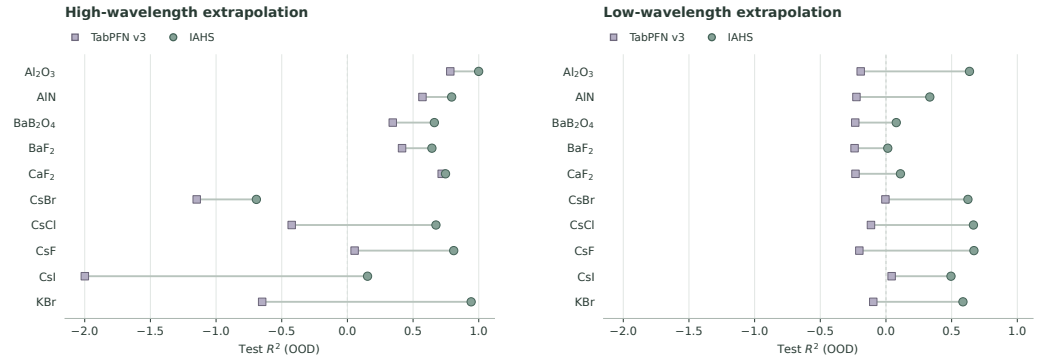


Figure 8: **Case-level optical OOD performance in both directions.** Left: high-wavelength extrapolation. Right: low-wavelength extrapolation. Each segment connects TabPFN v3 and IAHS on the same case. Both panels show all ten optical datasets in the same order with identical, unclipped linear axes. Mean R^2 is -0.1331 versus 0.5731 for high-wavelength extrapolation and -0.1497 versus 0.4213 for low-wavelength extrapolation, for TabPFN v3 and IAHS respectively. Segments represent paired score differences rather than uncertainty intervals.

Table 12: **Reference-based IAHS across three recorded seeds on the additional-benchmark suite.** Each row contains the same 41 non-OOD and 28 OOD evaluation cases.

Seed	Non-OOD mean	Non-OOD median	OOD mean	OOD median
42	0.8870	0.9512	0.2474	0.6396
43	0.8878	0.9499	0.2082	0.6337
44	0.8867	0.9492	0.1001	0.6300