NET WORK LISER

# Discussion Paper Series

IZA DP No. 19106

November 2026

# Variable Importance Recovery in Tabular Foundation Models: A Simulation Study

**Vasil Yasenov**
Adobe Inc., Stanford University
and IZA@LISER

**Nicholas Adams-Cohen**
Adobe Inc.

The IZA Discussion Paper Series (ISSN: 2365-9793) ("Series") is the primary platform for disseminating research produced within the framework of the IZA@LISER Network, an unincorporated international network of labour economists coordinated by the Luxembourg Institute of Socio-Economic Research (LISER). The Series is operated by LISER, a Luxembourg public establishment (établissement public) registered with the Luxembourg Business Registers under number J57, with its registered office at 11, Porte des Sciences, 4366 Esch-sur-Alzette, Grand Duchy of Luxembourg.

Any opinions expressed in this Series are solely those of the author(s). LISER accepts no responsibility or liability for the content of the contributions published herein. LISER adheres to the European Code of Conduct for Research Integrity. Contributions published in this Series present preliminary work intended to foster academic debate. They may be revised, are not definitive, and should be cited accordingly. Copyright remains with the author(s) unless otherwise indicated.

IZA DP No. 19106 November 2026

# Variable Importance Recovery in Tabular Foundation Models: A Simulation Study*


## Abstract

In machine learning, identifying which variables contribute predictive information can be as important as predictive accuracy. Tabular foundation models have matched or exceeded the predictive performance of tuned gradient boosting trees on many small- and medium-sized tabular datasets. Yet their ability to recover population-level variable importance has received little attention. We design and conduct a Monte Carlo simulation study to compare variable importance recovery across tabular foundation models and established machine learning baselines. Our framework distinguishes multiple population estimands, pairs each one with an aligned model-agnostic importance estimator, and evaluates recovery in terms of ranking, support, and relative magnitude. Our results show that tabular foundation models are generally competitive with established machine-learning baselines in recovering population variable importance, and often more accurate. We thus extend the evidence for their capabilities to a previously underexamined dimension, further strengthening their appeal.



## JEL classification

C45, C52, C15

## Keywords

variable importance, tabular foundation models, interpretability, Shapley values, simulation study, gradient boosting



## Corresponding author

Vasil Yasenov
yasenov@stanford.edu


* All errors are our own.

# 1 Introduction

A recent class of predictive methods, known as tabular foundation models (TFMs), uses labeled observations as context to form predictions without task-specific parameter estimation. On many small- and medium-sized tabular datasets, these methods match or outperform the predictive accuracy of gradient-boosted decision trees (GBDTs) without task-specific hyperparameter tuning (Hollmann et al. 2025; Grinsztajn et al. 2026; Qu et al. 2026; Ma et al. 2025; Kong and Das 2026). Yet predictive accuracy is not the only consideration in many applied settings. Policymakers and other decision-makers may also need to understand and defend a model's reliance on its inputs, and in some contexts may be willing to accept lower predictive performance in exchange for greater interpretability. One important component of this objective is identifying the variables that are most predictive of an outcome and quantifying their relative importance. Whether TFMs can accurately recover such importance remains largely unknown.

Existing work applies SHAP to TFMs and develops pretrained attribution models, expanding the tools available for assessing feature contributions (Hollmann et al. 2025; Fonseca and Stoyanovich 2026; Sena and Azevedo 2026). Related studies evaluate how faithfully attribution methods explain a fitted model (Agarwal et al. 2022; Li et al. 2023). An attribution may faithfully describe a fitted model's reliance on a variable without accurately reflecting that variable's predictive value in the population. These contributions advance attribution methodology, but do not establish how TFMs compare with established learners in recovering population-level variable importance.

This paper compares TFMs with established machine learning predictive baselines in their ability to recover population level variable importance. We design and conduct a Monte Carlo simulation study with known data-generating processes and evaluate recovery against population level ground truth. Our framework distinguishes four distinct population estimands from the importance of a fitted procedure and its finite-sample estimate. We pair each population target with an aligned model-agnostic importance estimator and evaluate importance recovery in terms of ranking, support, and relative magnitude. More broadly, this paper provides a framework for evaluating population variable importance recovery across AI-based prediction methods.

Our simulations show that TFMs are generally competitive with established machine-learning baselines in recovering population variable importance and often perform better. The size and consistency of these advantages depend on the definition of importance and the simulation setting, with some exceptions. The clearest gains arise when importance measures a variable's predictive value on its own: both TFMs outperform the boosted-tree baselines in identifying important variables and recovering their rankings and relative contributions. When importance measures the additional predictive value beyond the other inputs, TFMs also improve average ranking and magnitude recovery, although identification of the important variables is more often comparable. These findings complement the evidence on predictive accuracy, suggesting that strong TFM predictions need not come at the expense of reliable variable-importance recovery. Overall, they strengthen the case for considering TFMs.

The rest of the paper is organized as follows. Section 2 reviews the related literature and positions our study in context. Section 3 develops the conceptual framework, defining the population importance targets, their corresponding estimators, and the recovery metrics.

Section 4 describes the simulation design, and Section 5 presents the results. Section 6 discusses implications for model choice, limitations, and directions for future research. The appendices provide implementation details, numerical results, and additional diagnostics.

## 2 Literature Review

Research on TFMs increasingly extends from prediction to explanation. Predictive benchmarks compare TFMs with GBDTs across many datasets (Qu et al. 2025, 2026; Erickson et al. 2025). Interpretability studies also compare TFM explanations with classical learners, using SHAP (Lundberg and Lee 2017) or attention-based importance on applied datasets and synthetic data with known informative variables (Hollmann et al. 2025; Kolberg et al. 2025). Other work uses in-context learning for Shapley and LOCO computation, including conditional Shapley estimation (Rundel et al. 2024; Olsen and Christensen 2026). Pretrained attribution models estimate explanations from data without access to a fitted predictor (Fonseca and Stoyanovich 2026) or jointly produce predictions and explanations (Sena and Azevedo 2026). Together, these approaches expand the evidence and tools for interpreting TFMs. Evaluating population recovery additionally requires an explicit importance target and an appropriate evaluation criterion.

Variable importance is not a single quantity: different definitions capture different aspects of predictive information, particularly when features are dependent. Population-level frameworks define algorithm-agnostic importance through optimal predictive performance (Williamson and Feng 2020; Williamson et al. 2021, 2023). LOCO measures additional predictive value beyond the remaining covariates (Lei et al. 2018; Verdinelli and Wasserman 2024a), whereas Shapley effects allocate explained variance (Song, Nelson, and Staum 2016); feature dependence affects their interpretation (Verdinelli and Wasserman 2024b). Work on observational and interventional Shapley explanations, removal-based methods, and marginal versus conditional associations further clarifies these choices (Chen et al. 2020; Covert, Lundberg, and Lee 2020, 2021; Ewald et al. 2024). Permutation importance instead measures fitted-model reliance (Fisher, Rudin, and Dominici 2019), and unrestricted permutation can force extrapolation by breaking feature dependencies (Hooker, Mentch, and Zhou 2021). These distinctions make estimator–target alignment essential when comparing recovery across predictive procedures.

Simulation benchmarks offer complementary evaluation tools, ranging from synthetic data with computable reference quantities to broader assessments of attribution faithfulness (Liu et al. 2021; Agarwal et al. 2022; Li et al. 2023). Controlled local-attribution benchmarks also evaluate rank recovery and noise-feature mass against analytically or numerically derived reference attributions (Katakam 2026). Identifying designated informative variables, explaining a fitted model, and recovering a population importance quantity remain distinct evaluation objectives.

Our study combines these ideas to compare population variable-importance recovery across TFMs, gradient-boosted trees, and elastic net. We hold the importance procedures fixed across predictive models, match each to its population target, and evaluate support, rank, and relative magnitude recovery under known data-generating processes. This design shows how model comparisons depend on the importance definition, feature dependence, response function, and recovery criterion.

# 3 Conceptual Framework

## 3.1 Notation

Let $X = (X_1, \ldots, X_p)$ denote a vector of covariates and $Y$ an outcome, with $(X, Y) \sim P$. We write $[p] = \{1, \ldots, p\}$, $X_{-j}$ for all covariates except $X_j$, and $X_S$ for the subvector indexed by $S \subseteq [p]$. An i.i.d. training sample of size $n$ is denoted by $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$.

For a prediction function $f$ based on $X_S$, define its population risk under loss $\ell$ as

$$\mathcal{R}_S(f) = \mathbb{E}_P[\ell\{Y, f(X_S)\}] .$$

The corresponding population-optimal prediction function is

$$f_S^* \in \arg\min_f \mathcal{R}_S(f),$$

and we write $f^* = f_{[p]}^*$. Our simulations focus on regression, but the framework can also be adapted to classification. Under squared-error loss, $f_S^*(x_S) = \mathbb{E}[Y \mid X_S = x_S]$.

The data-generating process can equivalently be written as

$$Y = f^*(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid X] = 0.$$

Let $\widehat{f}$ denote the fitted model obtained by applying the chosen prediction procedure to the training sample.

The empirical risk of a fitted model $f$ based on $X_S$ is evaluated on a sample of size $m$ from $P$, independent of the training sample. Throughout, fitted models are held fixed when taking population expectations, and empirical prediction losses are computed on the independent evaluation sample. Reusing $(X_i, Y_i)$ for the evaluation observations only in the following expression, define

$$\widehat{\mathcal{R}}_S(f) = \frac{1}{m} \sum_{i=1}^m \ell\{Y_i, f(X_{i,S})\} .$$

For $S \subseteq [p]$, write $\mathcal{R}_S^* = \mathcal{R}_S(f_S^*)$ for the optimal population prediction error using $X_S$. Let $\mathcal{R}_\varnothing^*$ denote the prediction error of the optimal constant predictor. Define the predictive value of $X_S$ as

$$v(S) = \mathcal{R}_\varnothing^* - \mathcal{R}_S^*. \tag{1}$$

We distinguish three conceptual variable importance quantities:

- $\psi_j(P)$ denotes the *population importance* of variable $j$, a functional of the data-generating distribution that serves as the target estimand.
- $\phi_j$ denotes the corresponding *importance induced by fitted models under* $P$. It depends on $P$ and the fitted model or collection of fitted models used by the estimator.
- $\widehat{\phi}_j$ denotes the finite-sample estimate of $\phi_j$, computed using the independent evaluation sample and, where applicable, numerical approximation.

Thus, $\psi_j$ is the population target, $\phi_j$ is the importance induced by the fitted procedure when evaluated under $P$, and $\widehat{\phi}_j$ is its empirical estimate.

Let $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_p)$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$, and $\widehat{\boldsymbol{\phi}} = (\widehat{\phi}_1, \ldots, \widehat{\phi}_p)$. The total recovery error can then be decomposed as

$$\boldsymbol{\psi} - \widehat{\boldsymbol{\phi}} = \underbrace{(\boldsymbol{\psi} - \boldsymbol{\phi})}_{\text{model error}} + \underbrace{\left(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}\right)}_{\text{estimation error}} .$$

Model error is the discrepancy between the population importance target and the importance induced by the fitted procedure. It reflects both limitations of the prediction procedure and error from fitting on a finite training sample, rather than model misspecification alone. For example, a linear model like elastic net may fail to capture important nonlinear relationships in the population. Estimation error arises when evaluating the fitted procedure's importance using a finite independent evaluation sample and, where applicable, numerical approximation; it does not include training-sample error. Our study evaluates total recovery error without separately quantifying its model-error and estimation-error components. The next subsection defines the population importance targets we consider.

## 3.2 Population Level Importance *Estimands*, $\psi_j$

Population level importance, $\psi_j(P)$, describes the predictive information associated with $X_j$ under the joint distribution of $(X, Y)$ (Williamson et al. 2021, 2023). The question it answers is: *What value does $X_j$ provide for predicting $Y$ under $P$?* The answer depends on the definition of importance. We consider one binary and three continuous importance estimands.

### 3.2.1 Structural Importance

Structural importance records whether the population conditional outcome function genuinely depends on $X_j$ over the support of $P$:

$$\psi_j^{\text{str}} = \begin{cases} 1, & \text{if no function of } X_{-j} \text{ equals } f^*(X) \ P\text{-almost surely,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

It defines the active set $\{j : \psi_j^{\text{str}} = 1\}$ but provides no measure of magnitude. Because the definition is restricted to the support of $P$, variation at covariate combinations that cannot occur does not make a variable structurally important.

### 3.2.2 Marginal Importance

Marginal importance measures how much predictive performance improves when $X_j$ is used compared to a model with a constant only:

$$\psi_j^{\text{marg}} = v(\{j\}) = \mathcal{R}^*_{\varnothing} - \mathcal{R}^*_{\{j\}}. \tag{3}$$

Marginal importance can assign substantial importance to a variable that is merely correlated with the underlying signal, even if it provides no additional information after the other covariates are observed.

### 3.2.3 Conditional Importance

Conditional importance measures the additional predictive value of $X_j$ after all other covariates have been observed:

$$\psi_j^{\text{cond}} = v([p]) - v([p] \setminus \{j\}) = \mathcal{R}^*_{[p]\setminus\{j\}} - \mathcal{R}^*_{[p]} \geq 0. \tag{4}$$

This is the population target associated with leave-one-covariate-out procedures. It measures the unique contribution of $X_j$ and may therefore be small or zero when other covariates contain redundant predictive information.

### 3.2.4 Shapley Importance

Shapley importance averages the incremental predictive value of $X_j$ over all possible subsets of the remaining features (Williamson and Feng 2020):

$$\psi_j^{\text{Shap}} = \sum_{S\subseteq[p]\setminus\{j\}} \frac{|S|!(p-|S|-1)!}{p!} \left[v(S\cup\{j\}) - v(S)\right]. \tag{5}$$

Unlike conditional importance, which evaluates the contribution of $X_j$ only after all other covariates have entered, Shapley importance averages its contribution across every possible ordering. It therefore distributes shared predictive information among correlated or redundant variables. Under squared-error loss, $v(S) = \text{Var}(\mathbb{E}[Y \mid X_S])$, so $\psi_j^{\text{Shap}}$ corresponds to the unnormalized population Shapley effect (Song et al. 2016).

### 3.2.5 Discussion

The four definitions in Equations (2)–(5) measure distinct quantities and need not agree. Marginal importance credits correlated proxies, conditional importance measures unique information, Shapley importance distributes shared information, and structural importance records only whether a variable is required. Even with independent covariates, interactions can produce different rankings. Under squared loss, the continuous estimands coincide only in very special cases such as additive models with independent covariates.

For a simple example, let $Z, W, U$ be independent standard normal variables, let $Q = \rho Z + \sqrt{1-\rho^2}\,U$ be a proxy with $|\rho| < 1$, and let $Y = ZW + \varepsilon$ with independent mean-zero noise. Among the observed covariates $(Z, W, Q)$, only $Z$ and $W$ are structurally important, while all three marginal importance values are zero. Their conditional importance values are $1-\rho^2$, 1, and 0, respectively. Shapley importance can nevertheless allocate positive credit to $Q$ when $\rho \neq 0$, because it helps predict the interaction when $W$, but not $Z$, is observed. Thus, recovering the interaction's active variables and recovering marginal importance are different tasks, and assigning shared Shapley credit to a proxy is not an error.

The relevant estimand depends on the decision the analysis is intended to inform. For instance, structural importance suits settings in which the primary question is whether a variable belongs in the predictive signal at all, irrespective of the magnitude of its incremental contribution. Marginal importance is instead useful for screening variables, assessing the predictive content of a low-cost measurement on its own, or identifying variables that can serve as proxies when the full feature set is unavailable. Conditional importance is more

appropriate for incremental measurement or data-acquisition decisions. Shapley importance is preferable when correlated measurements jointly encode the same construct and credit must be allocated across them.

## 3.3 Fitted-Procedure Importance $\phi_j$

Let $\widehat{f}_S$ denote the model fitted on the training sample using only $X_S$, with $\widehat{f}_{[p]} = \widehat{f}$. We define the importance induced by these fitted models, $\phi_j$, using population expectations. The corresponding empirical estimates, $\widehat{\phi}_j$, replace these expectations with averages over an independent evaluation sample and, where needed, numerical approximation. Empirical risks are defined in Section 3.1.

We use model-agnostic variable importance estimators that can be applied to both TFMs and classical machine learning baselines. Tree-specific methods and scores, such as TreeSHAP, split gain, and Gini importance, are excluded because they are not applicable across all procedures.

### 3.3.1 Structural Importance

To evaluate structural importance recovery, we rank variables using each of the three importance estimators described below. We select the top $k$ variables, where $k$ is the known number of structurally active variables in the simulation, and compare the selected set with the true active set.

### 3.3.2 Marginal Refitting Importance

Marginal refitting importance compares a constant-only fit with a fit based on $X_j$ alone. The importance induced by these fitted models under $P$ is

$$\phi_j^{\text{marg}} = \mathcal{R}_{\varnothing}(\widehat{f}_{\varnothing}) - \mathcal{R}_{\{j\}}(\widehat{f}_{\{j\}}). \tag{6}$$

Here $\widehat{f}_{\varnothing}$ is the training-sample mean. Under squared-error loss, if the fitted models converge in mean square under $P$ to their corresponding population-optimal prediction functions, then $\phi_j^{\text{marg}}$ converges to $\psi_j^{\text{marg}}$.

### 3.3.3 Leave-One-Covariate-Out (LOCO)

Leave-one-covariate-out (LOCO) measures the change in population prediction error when $X_j$ is excluded and the prediction procedure is reapplied (Lei et al. 2018; Verdinelli and Wasserman 2024a). The importance induced by these fitted models under $P$ is

$$\phi_j^{\text{LOCO}} = \mathcal{R}_{[p]\setminus\{j\}}(\widehat{f}_{[p]\setminus\{j\}}) - \mathcal{R}_{[p]}(\widehat{f}). \tag{7}$$

The corresponding population target is $\psi_j^{\text{cond}}$. Unlike SHAP and SAGE, LOCO does not hold the fitted model fixed. It compares fitted models obtained with and without $X_j$. Under squared-error loss, if both fitted models converge in mean square under $P$ to their corresponding population-optimal prediction functions, then $\phi_j^{\text{LOCO}}$ converges to $\psi_j^{\text{cond}}$.

### 3.3.4 SAGE

SAGE assigns Shapley values to reductions in expected prediction loss rather than to individual predictions (Covert et al. 2020). Throughout, SAGE refers to its conditional version, which integrates missing features using their distribution conditional on the observed features. SAGE holds the full fitted model fixed; no new models are fitted for the feature subsets. We define its projection onto $X_S$ as $\widehat{f}_S^{\text{proj}}(x_S) = \mathbb{E}_P[\widehat{f}(X) \mid X_S = x_S]$.

The predictive value of this projected prediction function is

$$v_{\widehat{f}}(S) = \mathcal{R}_{\varnothing}(\widehat{f}_{\varnothing}^{\text{proj}}) - \mathcal{R}_S(\widehat{f}_S^{\text{proj}}),$$

where $\widehat{f}_{\varnothing}^{\text{proj}} = \mathbb{E}_P[\widehat{f}(X)]$ is the corresponding constant predictor. The SAGE importance under $P$ is

$$\phi_j^{\text{SAGE}} = \sum_{S \subseteq [p] \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \left[ v_{\widehat{f}}(S \cup \{j\}) - v_{\widehat{f}}(S) \right]. \tag{8}$$

Under squared-error loss, if the full fitted model converges to $f^*$ in mean square under $P$, then $\phi_j^{\text{SAGE}}$ converges to $\psi_j^{\text{Shap}}$.

Its empirical estimate $\widehat{\phi}_j^{\text{SAGE}}$ uses the independent evaluation sample and numerical approximation over coalitions and conditional feature distributions.

### 3.3.5 Discussion

Marginal refitting and LOCO compare models fitted on different feature sets; SAGE allocates the predictive performance of the full fitted model. Their paired population targets are marginal, conditional, and Shapley importance, respectively.

All procedures use $n$ observations for fitting and a separate, independently generated sample to evaluate prediction losses and variable importance. For TFMs, fitting supplies labeled context without updating pretrained weights; LOCO removes $X_j$ from both the context and the evaluation inputs. For tunable procedures, we select hyperparameters once per replication on the full feature set and reuse them when refitting restricted models, avoiding the computational cost of a separate cross-validated search for each feature subset. This protocol holds the selected settings fixed across refits, but settings appropriate for the full model need not be appropriate for every restricted prediction problem. Appendix A.3 gives the tuning protocol.

## 3.4 Evaluating Importance Recovery, $\psi_j - \widehat{\phi}_j$

We evaluate importance recovery through the discrepancy $\boldsymbol{\psi} - \widehat{\boldsymbol{\phi}}$, pairing each population target with its designated estimator. These pairings match the population definition, loss function, and conditioning structure. For each Monte Carlo replicate, we evaluate recovery along three complementary dimensions: support ($F_1$ score), rank (Spearman correlation), and magnitude (normalized RMSE). These dimensions capture different aspects of recovery: support asks whether the right variables were selected, rank assesses their ordering, and

magnitude assesses their relative sizes. Throughout, $\mathcal{I}$ denotes the evaluated covariates: all $p$ variables except under the restricted LOCO protocol of Appendix A.4. We describe each dimension and its evaluation metric in turn.

### 3.4.1 Support Recovery ($F_1$ Score)

Support recovery compares the top-ranked variables with a target active set. For marginal, conditional, and Shapley importance, we count target values strictly above zero as active:

$$S_\psi = \{j \in \mathcal{I} : \psi_j > 0\}.$$

This definition includes all positive analytical target values, however small. It does not threshold estimated scores or alter the values used for rank and magnitude recovery.

We select the $k = |S_\psi|$ variables with the largest estimated importance values to form $\widehat{S}_k$, then compute

$$\mathrm{F}_1 = \frac{2\,|\widehat{S}_k \cap S_\psi|}{|\widehat{S}_k| + |S_\psi|}.$$

When $k > 0$, both sets have size $k$, so $F_1$ equals precision and recall. The target support size is known rather than estimated. When $S_\psi = \varnothing$, $F_1$ is undefined.

### 3.4.2 Rank Recovery (Spearman Correlation)

Rank recovery measures agreement in the ordering of importance values. We include all evaluated variables except those designated as null in $\mathcal{J} \subseteq \mathcal{I}$ (Section 4.2), retaining variables even when $\psi_j$ or $\widehat{\phi}_j$ is zero. We use Spearman correlation:

$$\mathrm{RankRec} = \mathrm{cor}\left[\mathrm{rank}\left(\widehat{\boldsymbol{\phi}}_{\mathcal{J}}\right), \mathrm{rank}\left(\boldsymbol{\psi}_{\mathcal{J}}\right)\right]. \tag{9}$$

Exact ties receive midranks in each vector. Small estimation differences that break population ties can reduce correlation even when unequal target values are correctly ordered. Rank recovery is undefined with fewer than two compared variables or when either vector is constant over $\mathcal{J}$.

### 3.4.3 Magnitude Recovery (RMSE)

Magnitude recovery compares relative importance using signed $L_1$ normalization over all evaluated variables $\mathcal{I}$, including nulls:

$$\widetilde{\phi}_j = \frac{\widehat{\phi}_j}{\sum_{k\in\mathcal{I}} |\widehat{\phi}_k|}, \qquad \widetilde{\psi}_j = \frac{\psi_j}{\sum_{k\in\mathcal{I}} |\psi_k|}.$$

Signs are retained so negative estimates are penalized. We compute

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{I}|}\sum_{j\in\mathcal{I}} (\widetilde{\phi}_j - \widetilde{\psi}_j)^2}.$$

When defined, RMSE penalizes zero/nonzero mismatches between $\psi_j$ and $\widehat{\phi}_j$. It is undefined when either vector is entirely zero; such evaluations are reported separately.

# 4 Simulation Design

We generate repeated training samples from controlled data-generating processes for which the population importance vector $\boldsymbol{\psi}$ is known analytically. This section specifies the covariate distribution and variable roles, the response functions, the scenario grid, the sample sizes, and the ground truth computation needed to reproduce the design. The design varies sample size, feature count, within-pair correlation, and the response function. Estimator settings, model protocol, realized Monte Carlo precision and other details are deferred to Appendix A.

## 4.1 Predictive Algorithms

We evaluate two tabular foundation models, TabPFN-3 and TabICLv2, against three gradient-boosted tree implementations, XGBoost, LightGBM, and CatBoost, and an elastic-net baseline. We selected the TFMs for their prominence in tabular prediction, reflected in their popularity and citation counts, and the classical baselines for their widespread use in practice. All six procedures are paired with the same model-agnostic importance estimators rather than their native importance scores.

Computational cost materially constrained the simulation design. Importance estimation requires repeated model fitting or prediction across feature subsets, making a broad Monte Carlo comparison particularly costly for TFMs. We therefore balanced scenario coverage against tuning depth, estimator coverage, and Monte Carlo precision: we limited tuning, reduced the TabPFN-3 ensemble, used 50 replications per scenario, evaluated SAGE only at $p = 10$, and restricted LOCO feature evaluation when $n = 200, p = 50$. These trade-offs mean that our results compare the specified computational protocols, rather than the best attainable performance of each model family. Furthermore, each boosted-tree procedure receives a random search over ten hyperparameter configurations with five-fold cross-validation, and TabPFN-3 uses a four-member ensemble rather than the default eight. Software, model settings, and tuning protocols are given in Appendix A.3.

## 4.2 Covariates and Variable Roles

Covariates are drawn as $X \sim N(0, \Sigma)$ with unit marginal variances, $\Sigma_{jj} = 1$. Every covariate receives one of four construction labels:

- *signal*: a base variable that enters $f^*$, directly or through its average with a redundant variable. These variables let us assess recovery of genuine response drivers.
- *null*: independent of every other covariate and absent from $f^*$. They test whether estimators avoid assigning importance to irrelevant variables.
- *redundant*: an additional variable paired with a signal; both enter $f^*$ through their average, so swapping them leaves $f^*$ unchanged. They test whether estimators distinguish shared predictive information from a variable's unique contribution.
- *proxy*: an additional variable paired with a signal but absent from $f^*$; it is correlated with the signal when $\rho > 0$. They test whether estimators distinguish predictive information carried through correlation from direct involvement in the response function.

The active grid starts with $p_{\text{sig}}$ base signal variables. The first $n_{\text{red}} = \lfloor p_{\text{sig}}/2 \rfloor$ form redundant pairs with additional covariates, and the next $n_{\text{proxy}} = \lfloor p_{\text{sig}}/2 \rfloor$ form signal–proxy pairs

with additional covariates. The pairs are disjoint. Both members of a redundant pair are structurally active; designating one as the base signal is only bookkeeping. The remaining $p_{\text{null}} = p - p_{\text{sig}} - n_{\text{red}} - n_{\text{proxy}}$ covariates are null. Role assignment is independently re-randomized across the covariate positions $[p]$ for every Monte Carlo replicate, so recovery metrics are not confounded with covariate position.

$\Sigma$ is block-diagonal with a $2 \times 2$ block $\left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ for each redundant pair $(j, j')$ and each signal–proxy pair $(j, k)$, and identity elsewhere; $\rho$ controls correlation within each pair. For a redundant pair, $f^*$ replaces the plain term in $X_j$ with the symmetric average $(X_j + X_{j'})/2$. When the pair is strongly correlated, dropping either variable leaves most of the pair's predictive content in the other (small $\psi^{\text{cond}}$ for both). Shapley importance splits the pair's contribution equally. LOCO and SAGE aim to recover these respective patterns. A proxy $X_k$ never enters $f^*$, so $\psi_k^{\text{cond}} = 0$ exactly, but it can have positive marginal and Shapley importance when correlated with signal. Positive marginal-refitting or SAGE scores for proxies can therefore be appropriate, whereas LOCO targets zero.

This taxonomy separates target-specific behavior: independent nulls have no population importance, proxies can carry standalone or shared predictive information without a unique conditional contribution, and redundant pairs can have small conditional importance despite positive structural and Shapley importance. These are differences between targets, not necessarily failures of recovery. Finite fitted estimates need not equal these population targets.

For rank recovery, $\mathcal{J}$ contains all evaluated non-null covariates, retaining proxies. Excluding nulls avoids having the correlation dominated by separation of their zero-importance block from the remaining variables, rather than ordering within the latter (Appendix A.5).

### 4.3 Response Functions

Let $j_1, \ldots, j_{p_{\text{sig}}}$ index the base signal variables (in the order role assignment fixes them), and write $\tilde{X}_k$ for $X_{j_k}$, or for $(X_{j_k} + X_{j'_k})/2$ if $j_k$ belongs to a redundant pair $(j_k, j'_k)$ under the scenario's role assignment.

We consider four response functions:

$$f^*_{\text{lin}}(x) = \sum_{k=1}^{p_{\text{sig}}} \beta_k \tilde{X}_k, \qquad f^*_{\text{thr}}(x) = \sum_{k=1}^{p_{\text{sig}}} \beta_k \tilde{X}_k \mathbf{1}\{\tilde{X}_k > 0\},$$

$$f^*_{\text{nl}}(x) = \sum_{k=1}^{p_{\text{sig}}} \beta_k \, g_{(k \bmod 4)}(\tilde{X}_k), \qquad f^*_{\text{int}}(x) = \sum_{m=1}^{\lfloor p_{\text{sig}}/2 \rfloor} \beta_m \, \tilde{X}_{2m-1} \tilde{X}_{2m},$$

where $g_0(x) = x$, $g_1(x) = x^2 - 1$, $g_2(x) = \sin(2x)$, and $g_3(x) = |x| - \sqrt{2/\pi}$ are centered for a standard normal argument. All scenarios use coefficients equal to 1, and a target signal-to-noise ratio of 4. An averaged redundant pair has variance $(1 + \rho)/2$, so the same constants need not center its transformed response; additive constants do not affect these importance targets.

We then have:

$$Y = f^*(X) + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2), \quad \varepsilon \perp X$$

For the pure interaction response, $\psi_j^{\text{marg}} = 0$ for every variable because each term multiplies independent, centered signal blocks. Signal variables nevertheless have positive conditional and Shapley importance, reflecting differences between targets rather than a failure of marginal importance estimators. Support, rank, and normalized magnitude recovery are therefore undefined for this marginal target. We instead report marginal leakage, MargLeak $= \sum_{j=1}^{p} |\widehat{\phi}_j^{\text{marg}}| / \operatorname{Var}(Y)$, for which zero is the population target and lower values are better. This diagnostic measures total absolute deviation from the zero marginal target relative to outcome variance. It includes both positive and negative estimates; negative estimates occur when a single-variable refit has higher evaluation loss than the fitted constant. It is not a probability or a share constrained to lie below one.

## 4.4 Scenario Grid

The grid crosses $n \in \{200, 1000\}$, $p \in \{10, 50\}$, $\rho \in \{0, 0.5, 0.9\}$, and the four response functions, giving 48 scenarios. Covariate-role counts are determined by $p$, not independently crossed. For each scenario we compute the relevant structural, marginal, conditional, and Shapley population targets. All scenarios above keep $n$ and $p$ within the published context limits of the two primary TFMs (TabPFN-3 and TabICLv2), so neither model needs to subsample rows to fit within context.

We completed $B = 50$ Monte Carlo replicates for all grid scenarios, reflecting the compute cost of repeated full-context TFM fits and refits. Within each replicate, an independent evaluation sample of size $m = 2000$ is drawn fresh from $P$ and used both to compute test-sample prediction error and to estimate importance, $\widehat{\boldsymbol{\phi}}$. The training sample $\mathcal{D}_n$ is redrawn independently each replicate. SAGE uses a random 128-point subset of this evaluation sample.

The primary evaluation metric for prediction is normalized excess risk $(\widehat{R} - \sigma^2) / \operatorname{Var}(Y)$, where $\widehat{R}$ is the mean squared prediction loss of the full fitted model on this evaluation sample and $\sigma^2$ is the noise variance. SAGE settings and Monte Carlo precision are documented in Appendix A.

## 4.5 Ground Truth Computation

$v(S) = \operatorname{Var}(\mathbb{E}[Y \mid X_S])$ is computed analytically using Gaussian conditional moments. The additive responses decompose into independent signal-block contributions; each interaction combines two independent blocks. Shapley additivity therefore reduces the population game to local terms involving at most four covariates, even at $p = 50$. Exact local coalition enumeration preserves known ties and zeros rather than ranking numerical integration noise. Independent nulls receive zero importance under all targets; proxies have zero conditional importance.

We use 40,000 auxiliary Monte Carlo draws to estimate signal and outcome variances for noise calibration and normalization; the importance targets are computed analytically. Appendix A.8 derives the Gaussian moment identities and exact local Shapley calculation, and describes calibration and validation.

# 5 Results

Figures 1–4 present structural, marginal, conditional, and Shapley importance recovery, respectively. Each keeps the response functions separate and averages over sample size $n$, feature count $p$, and feature correlation $\rho$: 12 equally weighted scenario means per response, each based on $B = 50$ replications. SAGE was evaluated only at $p = 10$, so its results average six $n \times \rho$ scenario means per response. Sections 5.1–5.4 describe these figures in turn.

Figures 5 and 6 instead pool response functions to compare each TFM with mean boosted-tree performance. Figure 5 varies one design factor and averages over the other two; Figure 6 retains each $(n, p, \rho)$ combination and averages only over responses. Both weight the four responses equally, except for marginal recovery, which excludes the interaction-only response because its target is identically zero. Section 5.5 discusses these comparisons and the predictive-performance results.

Appendix B provides numerical recovery results in Tables B1–B3 and predictive performance in Table B4. Appendix C provides alternative recovery diagnostics in Figures C1–C3 and predictive-performance plots in Figure C4. Structural recovery is reported only in the main figures.

## 5.1 Structural Importance

Figure 1 contains four vertically stacked panels for the linear, threshold, nonlinear, and interaction-only responses. The horizontal axis identifies the six predictive procedures, grouped as TFMs, boosted trees, and elastic net; the vertical axis measures support recovery by $F_1$, with higher values indicating better recovery. Within each procedure, circles, squares, and triangles show results for marginal refitting, LOCO, and SAGE. Each estimator selects the known number of structurally active variables, and its selected set is compared with the structural active set. The estimators appear together because structural importance is a binary membership target, with no continuous importance estimator or rank and magnitude measures of its own.

Mean $F_1$ scores are shown as the colored markers and medians as black crosses. Bars are $\pm 1.96$ between-scenario standard errors, describing variation across the grid rather than within-scenario Monte Carlo precision. Vertical scales differ across panels; marked axis breaks omit empty ranges without excluding procedures or truncating error bars. SAGE's $p = 10$ scope differs from the other estimators' two feature counts, so differences between estimator series can also reflect the scenarios included. The same conventions also apply to Figures 2–4.

Overall, Figure 1 indicates broadly comparable structural recovery for the TFMs and boosted trees. LOCO $F_1$ averages about 0.94–0.99 for linear and threshold responses and 0.79–0.83 for interaction-only responses. Marginal refitting performs well outside the interaction-only design, where its zero population target cannot identify the joint signal. SAGE also recovers structural support well at $p = 10$ (nonlinear mean $F_1$: 0.978 for TabPFN-3 and 0.984 for TabICLv2). The elastic net is competitive for linear and threshold responses but substantially weaker for nonlinear and interaction-only responses (mean LOCO $F_1$: 0.517 and 0.365, respectively).

## 5.2 Marginal Importance

Figure 2 compares marginal refitting estimates with $\psi^{\text{marg}}$. Rows show the linear, threshold, and nonlinear responses; columns show support, rank, and magnitude recovery. Higher $F_1$ and Spearman correlation and lower normalized RMSE indicate better recovery. Each procedure has one estimator series, and colors distinguish model families.

The bottom panel reports marginal leakage, $\sum_j |\widehat{\phi}_j^{\text{marg}}| / \operatorname{Var}(Y)$, for the interaction-only response. Its marginal target is zero for every variable, leaving ordinary support, rank, and normalized magnitude recovery undefined. Lower leakage means smaller absolute deviations from this zero target, not better identification of the variables that jointly predict the outcome.

Under the specified refitting protocol, Figure 2 shows that both TFMs have higher mean support and rank recovery and lower magnitude error than each boosted tree in all three nondegenerate responses (nonlinear mean $F_1$: 0.942 for TabPFN-3 and 0.912 for TabICLv2, versus 0.774–0.845 for boosted trees). TabPFN-3 generally has stronger recovery than TabICLv2, particularly in relative magnitude (nonlinear mean RMSE: 0.014 versus 0.019, compared with 0.053–0.098 for boosted trees). Interaction-only leakage is also lower for the TFMs, averaging about 0.07 for TabPFN-3 and 0.32 for TabICLv2, compared with 1.7–6.5 for the boosted trees. The elastic net remains strong for linear and threshold responses and has very low leakage, but its nonlinear recovery is substantially poorer.

Boosted-tree single-variable refits reuse hyperparameters and tree counts selected using all covariates, without further tuning or early stopping. This choice can increase overfitting in restricted fits and yield negative importance estimates. Some of the observed recovery and leakage differences may therefore reflect the refitting protocol; separately tuned or early-stopped restricted fits could change their magnitude.

## 5.3 Conditional Importance

Figure 3 evaluates LOCO against $\psi^{\text{cond}}$ for all four responses, using the same layout as Figure 2. At $n = 200, p = 50$, LOCO evaluates a reduced feature universe. Interaction-only elastic-net rank metrics are undefined in 288 of 600 replicate evaluations and normalized magnitude metrics in 262; its summaries use only the defined values.

The results show higher mean rank recovery and lower magnitude error for both TFMs than for each boosted tree in every response class (interaction-only mean RMSE: 0.027 for TabPFN-3 and 0.028 for TabICLv2, versus 0.044–0.058 for boosted trees). Rank differences are modest for linear and threshold responses and larger for nonlinear and interaction-only responses, where TFM means are about 0.81–0.82 and 0.73–0.75, respectively (versus 0.74–0.78 and 0.65–0.71 for boosted trees). Support recovery is more comparable, with some boosted-tree means slightly exceeding those of the TFMs. An exception occurs at $n = 1000, p = 50, \rho = 0.9$ in the interaction-only design, conditional $F_1$ is approximately 0.67 for TabPFN-3 and 0.57 for TabICLv2, versus 0.75 for the mean boosted-tree benchmark. The elastic net again performs well for linear and threshold responses but much less well for nonlinear and interaction-only responses.

## 5.4 Shapley Importance

Figure 4 evaluates SAGE against $\psi^{\text{Shap}}$, using the same layout as Figure 3. Results are restricted to $p = 10$ because SAGE was not evaluated at $p = 50$.

Recovery is broadly similar across the TFMs and boosted trees, rather than a consistent advantage for one procedure on every criterion (nonlinear mean Spearman correlation: 0.803 for TabPFN-3 and 0.812 for TabICLv2, versus 0.788–0.816 for boosted trees). Interaction-only support recovery is high for the TFMs, CatBoost, and LightGBM, with mean $F_1$ around 0.92–0.96. The clearest separation is from the elastic net, which has much poorer nonlinear and interaction-only recovery (mean $F_1$: 0.485 and 0.362, respectively). Both TFMs are competitive for Shapley recovery within the evaluated $p = 10$ scope. Small differences among the stronger procedures should be interpreted cautiously, since SAGE's finite evaluation sample, conditional integration, and permutation sampling contribute estimation error alongside fitted-model error.

## 5.5 Additional Results

Figure 5 presents seven rows: structural support $F_1$, followed by support $F_1$ and Spearman rank recovery for marginal, conditional, and Shapley importance. Columns vary sample size, feature count, and feature correlation, with the factor's levels on the horizontal axis. The vertical axis is the TFM's recovery metric minus the equally weighted mean metric for XGBoost, LightGBM, and CatBoost; positive values favor the TFM because higher $F_1$ and Spearman correlation indicate better recovery. Large points and lines average over responses and the two design factors not on the horizontal axis, while faint points show contrasts within design cells, averaged over responses. Shapley feature-count panels show unconnected points at $p = 10$, with $p = 50$ marked as not evaluated.

Figure 5 supports broad TFM competitiveness in these pooled comparisons. Marginal support and rank advantages are especially apparent, and larger at $p = 50$ (mean $F_1$ advantages of 0.169 for TabPFN-3 and 0.132 for TabICLv2); many structural, conditional-support, and Shapley comparisons are closer to zero. Small negative Shapley differences in some pooled views (e.g., a mean $F_1$ contrast of $-0.004$ for TabICLv2 at $n = 1000$) are better read as broadly similar performance than as substantial reversals. These averages do not establish superiority in every response, scenario, or comparison with an individual boosted tree.

Figure 6 retains each $(n, p, \rho)$ combination in four target-specific blocks. Rows identify design cells and columns identify the two TFMs. The structural block shows LOCO support $F_1$; the other blocks show Spearman rank recovery and normalized RMSE. Higher $F_1$ and Spearman correlation and lower raw RMSE indicate better recovery; larger plotted contrasts favor the TFM for every metric. Responses are pooled with the same weights and scope as Figure 5. Support and rank contrasts subtract mean boosted-tree performance from TFM performance; RMSE reverses the subtraction, so positive values always favor the TFM. Blue favors the TFM and orange the boosted-tree mean, with scales shared within each metric. Gray SAGE cells were not evaluated, and asterisks flag the $n = 200, p = 50$ restricted LOCO scope. Neither colors nor the points in Figure 5 indicate statistical significance.

Figure 6 reinforces the broad marginal-recovery advantage and generally favorable conditional rank and magnitude comparisons, while showing more mixed Shapley contrasts. Some marginal advantages are substantial: at $n = 200, p = 50, \rho = 0.5$, rank recovery exceeds the

mean boosted-tree benchmark by about 0.21 for TabPFN-3 and 0.18 for TabICLv2 (mean RMSE is lower by 0.032 and 0.023, respectively, in the same cell). Shapley differences are generally smaller, especially at $n = 1000$. Together, Figures 5 and 6 locate the gains without turning small cell-specific differences into a universal ranking.

Figures C1–C3 in Appendix C provide response-specific checks using off-support mass, Kendall's $\tau_b$, and normalized mean absolute deviation in place of $F_1$, Spearman correlation, and RMSE. They preserve the broad marginal-recovery advantage and more mixed conditional and Shapley comparisons, although individual method orderings can change. Tables B1–B3 in Appendix B give the numerical summaries for both sets of metrics and the counts of undefined evaluations.

Figure C4 and Table B4 report predictive performance. The figure has one panel per response function, with predictive procedures on the horizontal axis and normalized excess risk on the vertical axis; lower values are better. Points average the same 12 $n \times p \times \rho$ scenario means used for the non-SAGE recovery comparisons.

Figure C4 shows that both TFMs have lower mean normalized excess risk than each boosted tree in every response class; the elastic net is best for the linear response (mean normalized excess risk: 0.014, versus 0.019 for TabPFN-3 and 0.018 for TabICLv2). Interaction-only means are about 0.22 and 0.27 for the TFMs, versus 0.40–0.55 for boosted trees. Thus, strong predictive performance accompanies generally competitive importance recovery, although the predictive ordering does not transfer to every recovery criterion.

# 6 Discussion

This paper asks whether tabular foundation models can compete with established machine-learning baselines in recovering population variable importance. Our Monte Carlo framework makes this comparison systematic by distinguishing population targets from fitted-model estimates, matching targets and estimators, and evaluating support, rank, and relative magnitude recovery against known ground truth. The results show broad competitiveness, with clearer marginal-recovery advantages and more comparable performance on other criteria, although conditional support recovery is lower in a highly correlated interaction setting. Alongside their predictive performance, these findings show that the strengths of this new generation of models extend to an important, previously underexamined aspect of model interpretation.

These findings do not make TFMs the preferred choice for every application. Population variable-importance recovery does not measure human comprehensibility or transparency of the model itself. Requirements for an inspectable functional form, explicit coefficients, or a sparse specification such as elastic net can favor established models regardless of their importance-recovery performance. Computational cost, fitting and inference speed, scalability, tuning effort, and deployment requirements also matter, as do reproducibility, robustness, calibration, and uncertainty quantification. Established baselines may offer advantages on these dimensions, which our study does not benchmark. Model choice therefore depends on the combination of predictive performance, explanation capabilities, and operational requirements that fits the application.

Our study has important limitations. The evidence is restricted to continuous Gaussian covariates with controlled dependence and four regression response functions, rather than

the full diversity of applied tabular data. We evaluate two TFMs and three importance estimators under specific fitting and computational protocols, so the comparisons should not be interpreted as isolating architectural advantages or establishing optimal performance for each model family. Reusing full-model hyperparameters may disadvantage classical baselines when the appropriate model complexity changes with the feature set, particularly for single-variable boosted-tree refits that retain the full-model tree count without further early stopping. Some of the observed TFM advantages in marginal and conditional importance recovery may therefore reflect this protocol; our experiments do not establish how these comparisons would change under subset-specific tuning. SAGE uses finite evaluation data and Monte Carlo approximations with a drift-based stopping rule, which limits the resolution of small Shapley-recovery differences. Computational constraints also restrict the evaluated scope.

Future work should extend the simulations to include categorical and missing covariates, more complex dependence and interactions, heterogeneous noise, and distribution shifts, alongside larger datasets. Evaluating estimated conditional distributions and tuning performed separately for each feature subset would help separate model performance from explanation and fitting protocols. Using the known covariate distribution also leaves open how recovery changes when that distribution must be estimated. A distinct extension concerns causality: our targets describe predictive information, not effects of intervening on variables. Explicit causal structures could support evaluation of causally relevant variables or intervention-based quantities, but would require corresponding causal estimands and identification assumptions.

## AI Disclosure Statement

Generative AI tools (Claude, Copilot) assisted with literature review, analytical computation of population targets, code implementation, writing and revising the simulation design, producing an initial manuscript draft, and preparing the full supplementary appendix. The authors developed the research idea, the scientific and literature framing, the estimands, and the simulation analyses. They reviewed and revised all AI-assisted material, tested all code, verified the reported results against stored simulation outputs, and checked each reference against its source. The authors take full responsibility for the content, accuracy, and integrity of this paper.

# 7 References


Agarwal, Chirag, Satyapriya Krishna, Eshika Saxena, et al. 2022. "OpenXAI: Towards a Transparent Evaluation of Model Explanations." *Advances in Neural Information Processing Systems*.

Chen, Hugh, Joseph D. Janizek, Scott M. Lundberg, and Su-In Lee. 2020. "True to the Model or True to the Data?" *arXiv Preprint arXiv:2006.16234*.

Covert, Ian, Scott M. Lundberg, and Su-In Lee. 2020. "Understanding Global Feature Contributions with Additive Importance Measures." *Advances in Neural Information Processing Systems* 33.

Covert, Ian, Scott M. Lundberg, and Su-In Lee. 2021. "Explaining by Removing: A Unified Framework for Model Explanation." *Journal of Machine Learning Research* 22 (209): 1–90.

Erickson, Nick, Lennart Purucker, Andrej Tschalzev, et al. 2025. "TabArena: A Living Benchmark for Machine Learning on Tabular Data." *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Ewald, Fiona Katharina, Ludwig Bothmann, Marvin N. Wright, Bernd Bischl, Giuseppe Casalicchio, and Gunnar König. 2024. "A Guide to Feature Importance Methods for Scientific Inference." *World Conference on Explainable Artificial Intelligence*, 440–64. https://doi.org/10.1007/978-3-031-63797-1_22.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. 2019. "All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." *Journal of Machine Learning Research* 20 (177): 1–81.

Fonseca, João, and Julia Stoyanovich. 2026. "ExplainerPFN: Towards Tabular Foundation Models for Model-Free Zero-Shot Feature Importance Estimations." *arXiv Preprint arXiv:2601.23068*.

Grinsztajn, Léo, Klemens Flöge, Oscar Key, et al. 2026. "TabPFN-3: Technical Report." *arXiv Preprint arXiv:2605.13986*.

Hollmann, Noah, Samuel Müller, Lennart Purucker, et al. 2025. "Accurate Predictions on Small Data with a Tabular Foundation Model." *Nature* 637: 319–26. https://doi.org/10.1038/s41586-024-08328-6.

Hooker, Giles, Lucas Mentch, and Siyu Zhou. 2021. "Unrestricted Permutation Forces Extrapolation: Variable Importance Requires at Least One More Model, or There Is No Free Variable Importance." *Statistics and Computing* 31 (6): 82. https://doi.org/10.1007/s11222-021-10057-z.

Katakam, Raj Kiran Gupta. 2026. "AGOP-IxG: A Gradient Covariance Filter for Local

Feature Attribution on Tabular Data, with a Controlled Benchmark." *arXiv Preprint arXiv:2605.15700*.

Kolberg, Christopher, Jules Kreuer, Jonas Huurdeman, Sofiane Ouaari, Katharina Eggensperger, and Nico Pfeifer. 2025. "TabPFN-Wide: Continued Pre-Training for Extreme Feature Counts." *arXiv Preprint arXiv:2510.06162*.

Kong, Weihao, and Abhimanyu Das. 2026. *Introducing TabFM: A Zero-Shot Foundation Model for Tabular Data*. Google Research Blog. https://research.google/blog/introducing-tabfm-a-zero-shot-foundation-model-for-tabular-data/.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113 (523): 1094–111. https://doi.org/10.1080/01621459.2017.1307116.

Li, Xuhong, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. 2023. "$\mathcal{M}^4$: A Unified XAI Benchmark for Faithfulness Evaluation of Feature Attribution Methods Across Metrics, Modalities and Models." *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Liu, Yang, Sujay Khandagale, Colin White, and Willie Neiswanger. 2021. "Synthetic Benchmarks for Scientific Research in Explainable Machine Learning." *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Lundberg, Scott M., and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30.

Ma, Junwei, Valentin Thomas, Rasa Hosseinzadeh, et al. 2025. "TabDPT: Scaling Tabular Foundation Models on Real Data." *Advances in Neural Information Processing Systems* 38.

Olsen, Lars Henry Berge, and Dennis Christensen. 2026. "Computing Conditional Shapley Values Using Tabular Foundation Models." *arXiv Preprint arXiv:2602.09489*.

Qu, Jingang, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. 2025. "TabICL: A Tabular Foundation Model for in-Context Learning on Large Data." *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of machine learning research, vol. 267: 50817–47.

Qu, Jingang, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. 2026. "TabICLv2: A Better, Faster, Scalable, and Open Tabular Foundation Model." *Proceedings of the 43rd International Conference on Machine Learning*.

Rundel, David, Julius Kobialka, Constantin von Crailsheim, Matthias Feurer, Thomas Nagler, and David Rügamer. 2024. "Interpretable Machine Learning for TabPFN." *World Conference on Explainable Artificial Intelligence*, 465–76. https://doi.org/10.1007/978-3-031-63797-1_23.

Sena, Luan Borges Teodoro Reis, and Francisco Galuppo Azevedo. 2026. "Real-Time Explanations for Tabular Foundation Models." *DATA4Science Workshop at ICLR*.

Song, Eunhye, Barry L. Nelson, and Jeremy Staum. 2016. "Shapley Effects for Global Sensitivity Analysis: Theory and Computation." *SIAM/ASA Journal on Uncertainty Quantification* 4 (1): 1060–83. https://doi.org/10.1137/15M1048070.

Verdinelli, Isabella, and Larry Wasserman. 2024a. "Decorrelated Variable Importance." *Journal of Machine Learning Research* 25 (7): 1–27.

Verdinelli, Isabella, and Larry Wasserman. 2024b. "Feature Importance: A Closer Look at Shapley Values and LOCO." *Statistical Science* 39 (4): 623–36. https://doi.org/10.1214/24-STS937.

Williamson, Brian D., and Jean Feng. 2020. "Efficient Nonparametric Statistical Inference on Population Feature Importance Using Shapley Values." *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of machine learning research, vol. 119: 10282–91.

Williamson, Brian D., Peter B. Gilbert, Marco Carone, and Noah Simon. 2021. "Nonparametric Variable Importance Assessment Using Machine Learning Techniques." *Biometrics* 77 (1): 9–22. https://doi.org/10.1111/biom.13392.

Williamson, Brian D., Peter B. Gilbert, Noah R. Simon, and Marco Carone. 2023. "A General Framework for Inference on Algorithm-Agnostic Variable Importance." *Journal of the American Statistical Association* 118 (543): 1645–58. https://doi.org/10.1080/01621459.2021.2003200.

# 8 Figures

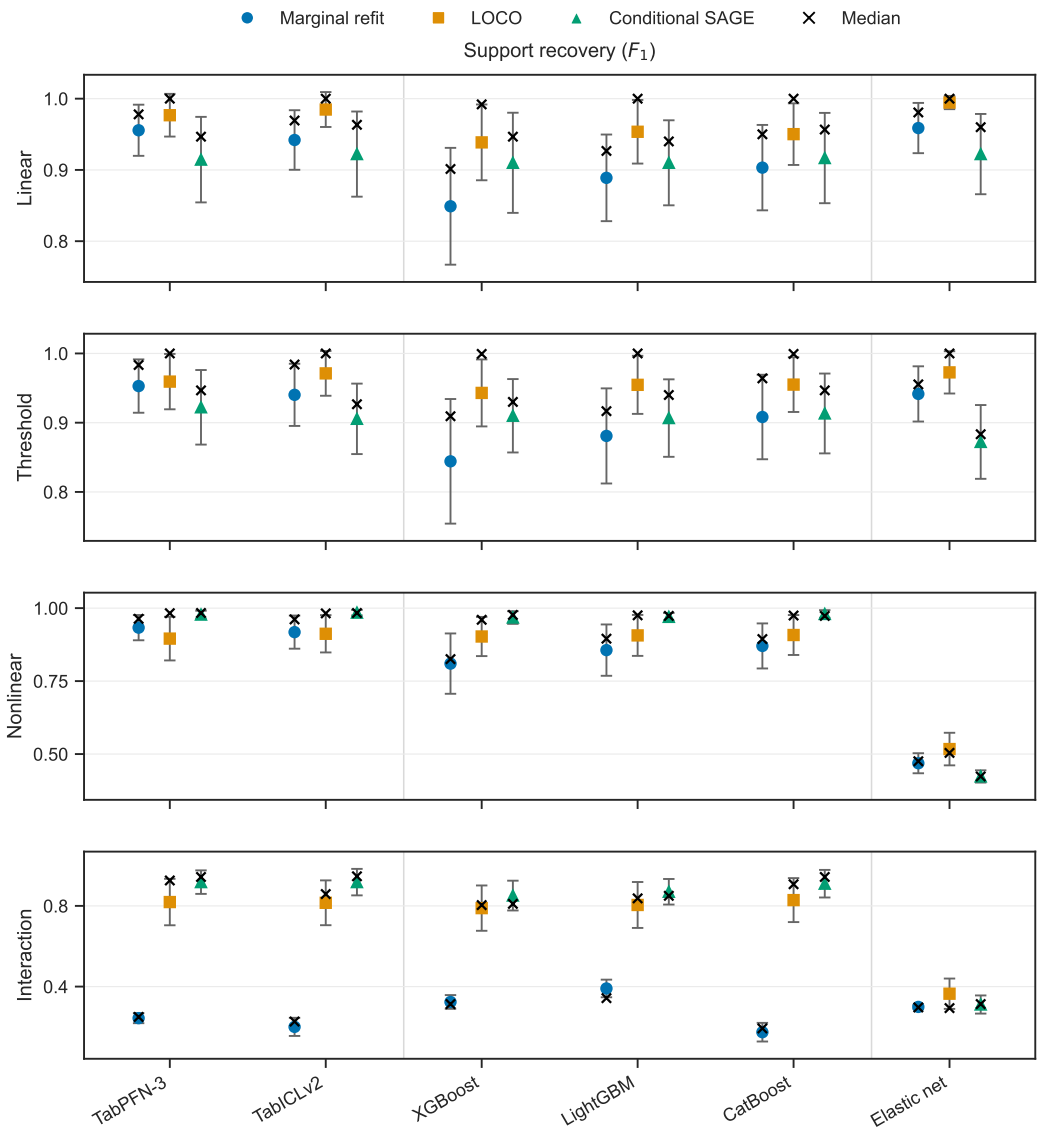



Figure 1: **Structural support recovery**. Marginal refitting and LOCO average over $n \times p \times \rho$ (12 scenarios per response function); SAGE averages over $n \times \rho$ at $p = 10$ (six scenarios). Points are the mean $F_1$ score of each estimator's selected set $\widehat{S}$ against the structural active set $S_{\psi^{\text{str}}} = \{j : \psi_j^{\text{str}} = 1\}$; error bars are $\pm 1.96$ between-scenario standard errors across each estimator's scenario means, and $\times$ marks their median.

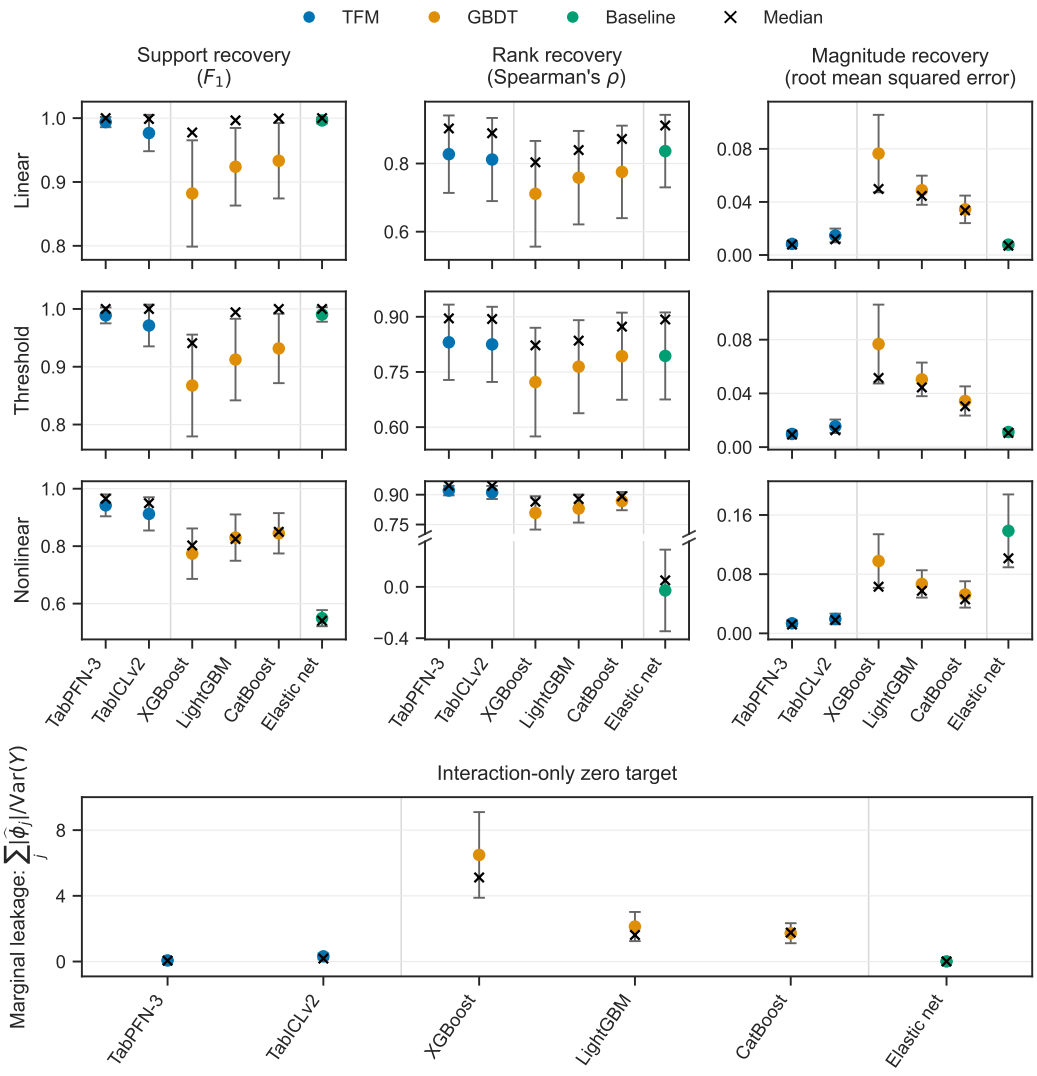



Figure 2: **Marginal importance recovery**. The upper grid reports $F_1$, Spearman $\rho$, and normalized RMSE for the three nondegenerate response functions, averaged over the 12 $n \times p \times \rho$ scenario means. The bottom panel reports marginal leakage for the interaction-only response. Points are means, error bars are $\pm 1.96$ between-scenario standard errors, and $\times$ marks the median.

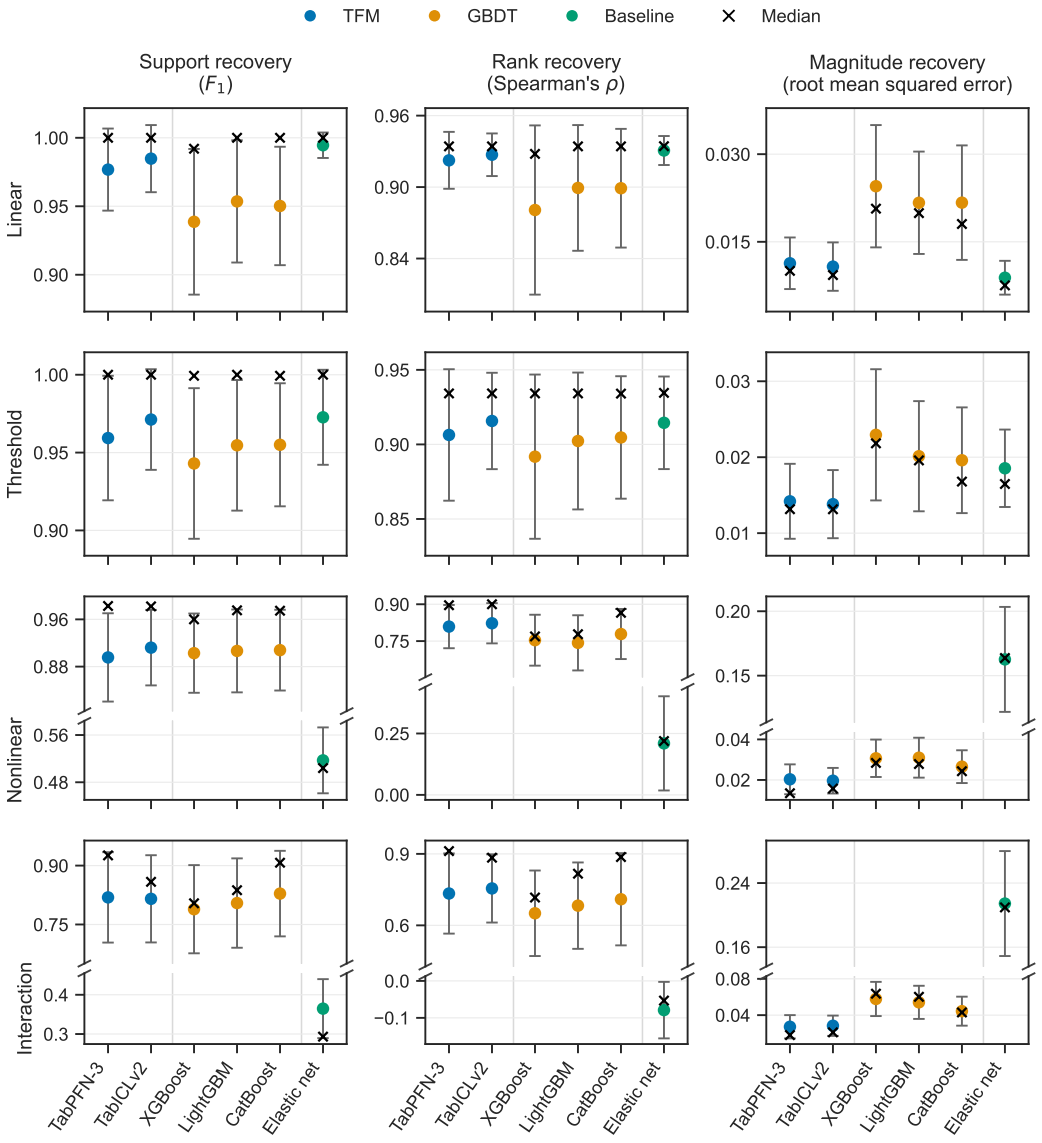



Figure 3: **Conditional importance recovery**. All metrics are averaged over the design grid ($n \times p \times \rho$). Rows are predictive functions; columns are $F_1$ (support recovery), Spearman $\rho$ (rank recovery), and normalized RMSE (magnitude recovery). Error bars are $\pm 1.96$ between-scenario standard errors, and $\times$ marks the median.

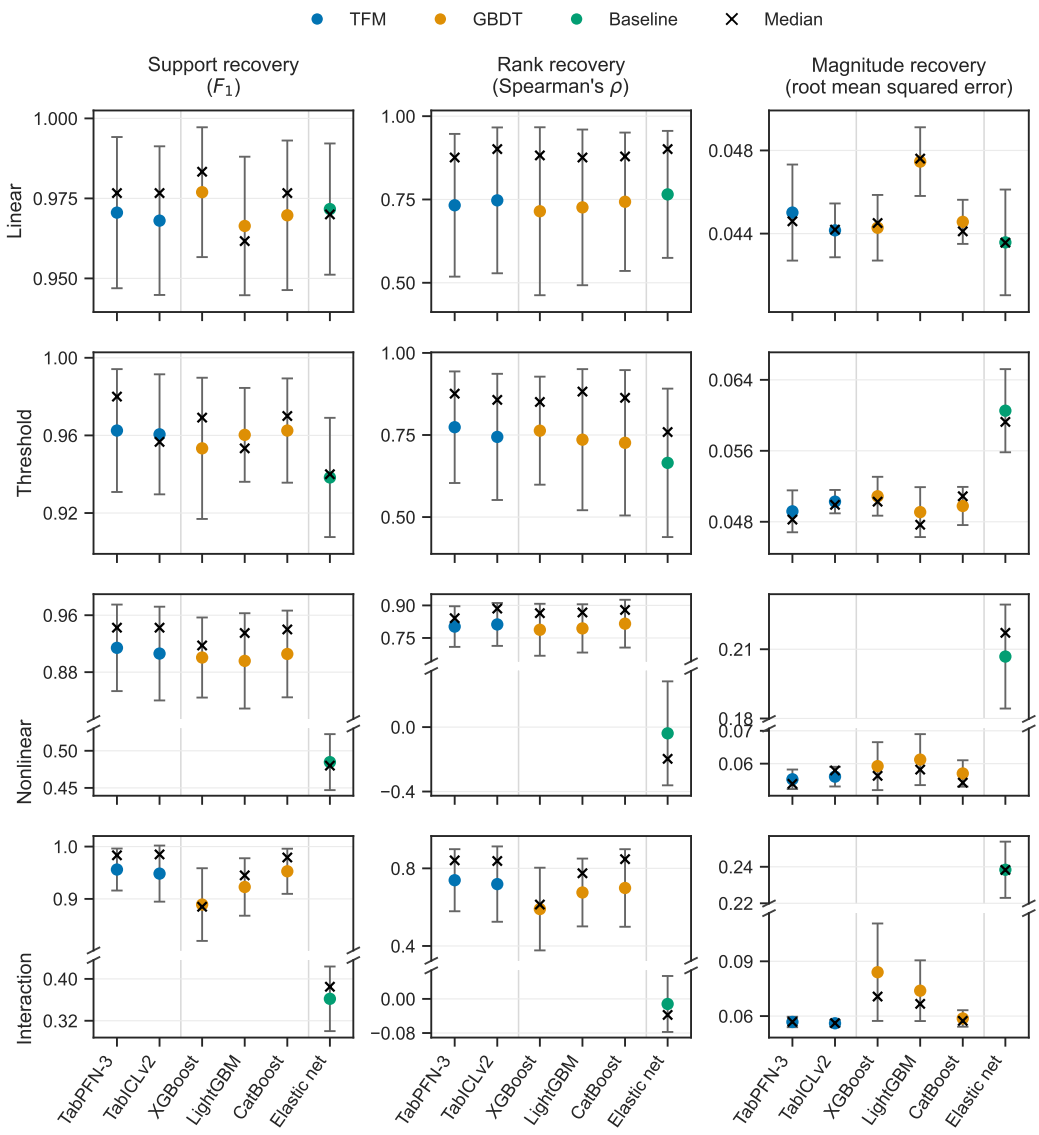



Figure 4: **Shapley importance recovery**. Results restricted to $p = 10$ and averaged over the six $n \times \rho$ scenario means. Rows are predictive functions; columns are $F_1$ (support recovery), Spearman $\rho$ (rank recovery), and normalized RMSE (magnitude recovery). Error bars are $\pm 1.96$ between-scenario standard errors, and $\times$ marks the median.

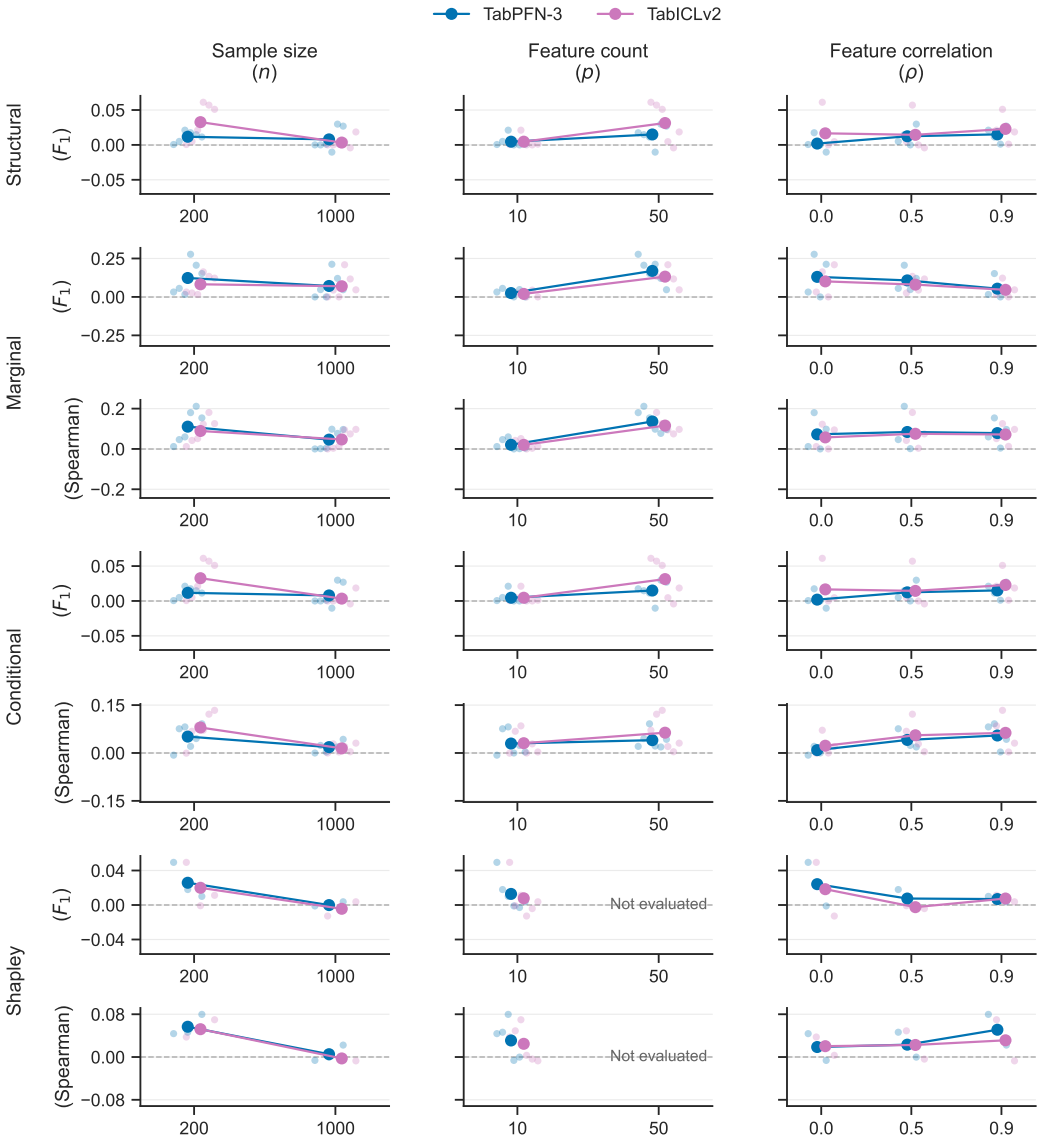



Figure 5: **Importance recovery averaged over response types.** All results are relative to mean XGBoost, LightGBM, and CatBoost performance ($B = 50$). Rows show matched-estimator support $F_1$ and Spearman rank recovery; structural recovery uses LOCO support only. Columns vary one design factor, averaging equally over the other two and over responses (four except marginal's three nondegenerate responses). Large points and lines show subgroup means; faint points show contrasts within design cells, averaged over responses.

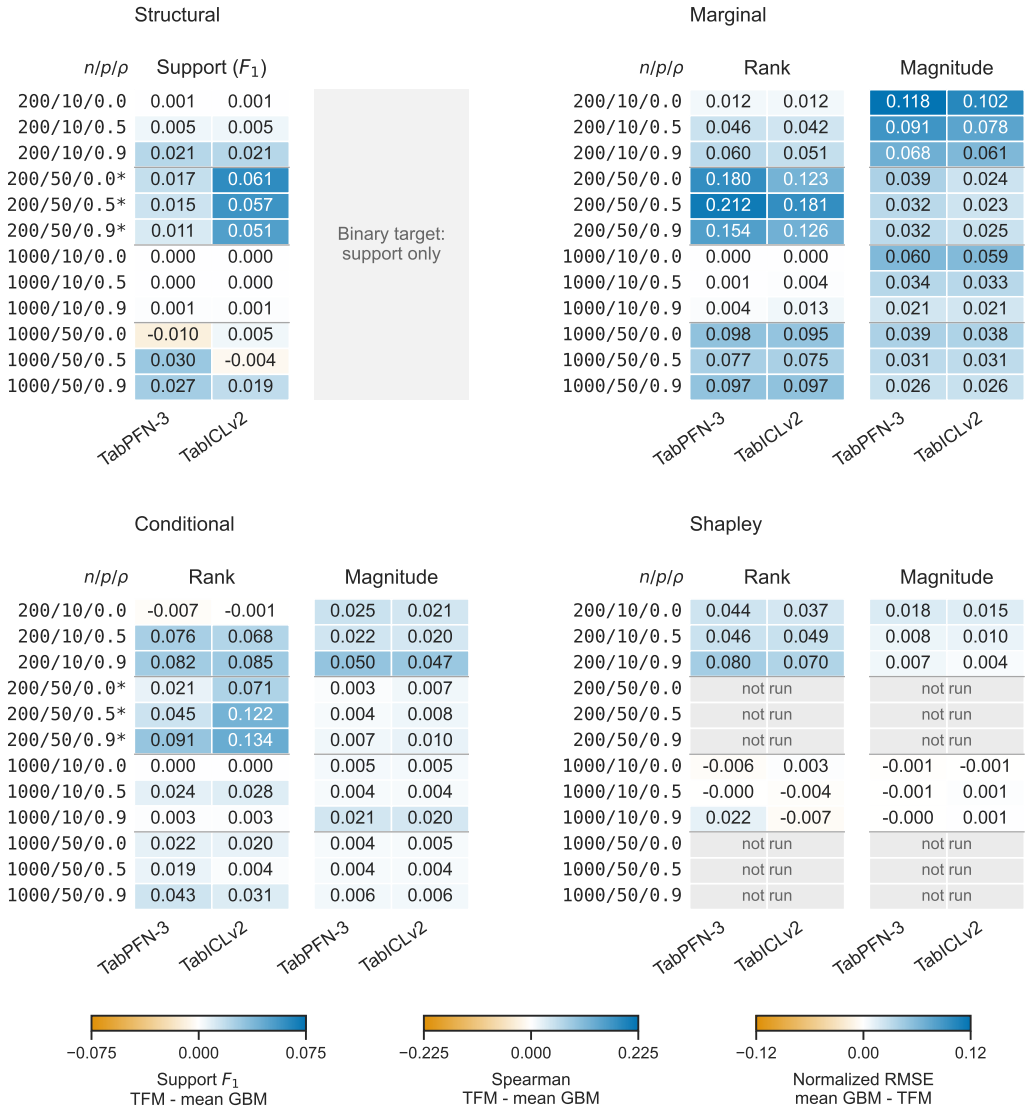



Figure 6: **Importance recovery within design cells.** All results are relative to mean XGBoost, LightGBM, and CatBoost performance ($B = 50$). Rows within each target identify $(n, p, \rho)$; columns identify the TFM. Structural recovery is LOCO-selected support $F_1$. Rank recovery uses Spearman correlation; magnitude recovery uses signed-$L_1$-normalized RMSE. Support and rank differences are TFM minus mean GBM; RMSE differences reverse the subtraction so positive values always favor the TFM. Blue favors the TFM and orange the GBM mean. Color scales differ by metric, not by target within a metric. Gray SAGE cells were not evaluated; the structural target has no rank or magnitude metric. An asterisk marks restricted LOCO scope at $n = 200, p = 50$. Colors encode point estimates, not statistical significance.