Context-Constrained Transfer Learning for Tabular Foundation Models via Data Distillation

Abstract

Tabular Foundation Models (TFMs) have demonstrated remarkable success as black-box inference engines, operating fundamentally through parameter-free adaptation via in-context learning. However, their utility in transfer learning is severely limited by strict context-size constraints and a high sensitivity to distribution shifts. These limitations frequently lead to negative transfer when heterogeneous source data is integrated. To overcome these dual bottlenecks, we propose Context-Constrained Transfer Learning via Data Distillation (TL-ANDI). Rather than indiscriminately pooling source data, TL-ANDI first identifies representative anchor points based on empirical assessments of both covariate and posterior shifts. Subsequently, it employs local kernel distillation to smooth labels, effectively reducing the variance of the source distribution while strictly adhering to the architectural limits of the TFM. We establish theoretical guarantees for this novel distillation mechanism. Extensive empirical evaluations on simulated setups and real-world datasets demonstrate that TL-ANDI robustly mitigates negative transfer and significantly outperforms naive transfer learning baselines in both regression and classification problems.

Keywords: transfer learning, tabular foundation model, covariate shift, posterior shift
Mathematics Subject Classification (2020): 62G05, 68T05

1 Introduction

Tabular Foundation Models (TFMs), such as TabPFN (Hollmann et al., 2025) and Limix (Zhang et al., 2025) have recently achieved remarkable success in tabular data analysis, largely driven by the paradigm of In-Context Learning (ICL). By taking inputs that include the training data $(X^{(\mathrm{tr})}, \boldsymbol{y}^{(\mathrm{tr})}) \in \mathbb{R}^{n_{\mathrm{tr}} \times (p+1)}$ and the test covariates matrix $X^{(\mathrm{test})} \in \mathbb{R}^{n_{\mathrm{test}} \times p}$ as context, a TFM acts as a black-box inference engine, directly outputting predictions $\hat{\boldsymbol{y}}^{(\mathrm{test})} \in \mathbb{R}^{n_{\mathrm{test}}}$.

Despite their empirical superiority on standard benchmarks, current TFMs are inherently restricted by their single-task modeling nature. They operate under the implicit assumption that the context data originates from a single, homogeneous distribution. In practice, however, it is often necessary to integrate knowledge from multiple heterogeneous source domains to improve predictive performance on a target task—a scenario that falls squarely into the realm of transfer learning (Torrey and Shavlik, 2010). To address this limitation, we investigate the adaptation of TFMs, specifically focusing on TabPFN, for transfer learning across heterogeneous tabular datasets.

Directly pooling source and target data into the TFM's context is usually infeasible due to two fundamental bottlenecks. First, from a statistical perspective, TFMs are highly sensitive

to distribution shift; incorporating source data with a differing distribution can cause negative transfer and severely degrade the model's accuracy on the target task. Second, from an architectural perspective, models like TabPFN are strictly bounded by their maximum context size, making it impossible to ingest large-scale source datasets.

To simultaneously overcome these dual challenges of distribution shift and bounded context length, we propose a novel framework: Context-Constrained Transfer Learning via Data Distillation. Instead of indiscriminately pooling data, our method strategically distills a highly informative, strictly bounded subset of source samples to augment the target context, thereby maximizing knowledge transfer while strictly adhering to the model's architectural limits.

1.1 Problem Formalization

Let $\mathcal{D}^{(\mathrm{src})} = \{\boldsymbol{x}_i^{(\mathrm{src})}, y_i^{(\mathrm{src})}\}_{i=1}^{n_{\mathrm{src}}}$ be a large-scale source dataset and $\mathcal{D}^{(\mathrm{tar})} = \{\boldsymbol{x}_i^{(\mathrm{tar})}, y_i^{(\mathrm{tar})}\}_{i=1}^{n_{\mathrm{tar}}}$ be a relatively small-scale target data. Our goal is to make predictions for test data which are sampled from the target population. For the test data, we observe test covariates $\{\boldsymbol{x}_i^{(\mathrm{test})}\}_{i=1}^{n_{\mathrm{test}}}$ at the training phase.

TFMs, denoted as the operator $\mathcal{P}$, generate predictive estimates for target queries based on a provided context $\mathcal{C}^{(\mathrm{tr})} \cup \mathcal{C}^{(\mathrm{test})}$:

$$\hat{y}^{(\mathrm{test})} = \mathcal{P}(\mathcal{C}^{(\mathrm{test})} \mid \mathcal{C}^{(\mathrm{tr})}, \mathcal{C}^{(\mathrm{test})}), \tag{1}$$

where $\mathcal{C}^{(\mathrm{tr})}$ denotes the labeled training data (training context) and $\mathcal{C}^{(\mathrm{test})}$ denotes the unlabeled test data (test context). For instance, in the single-task setting, where the source data is not available, we have $\mathcal{C}^{(\mathrm{tr})} = \mathcal{D}^{(\mathrm{tar})}$ and $\mathcal{C}^{(\mathrm{test})} = \{X^{(\mathrm{test})}\}$. With source data at hand, we will investigate how to summarize $\mathcal{D}^{(\mathrm{src})}$ and $\mathcal{D}^{(\mathrm{tar})}$ into the training context $\mathcal{C}^{(\mathrm{tr})}$.

Due to the architectural constraints of the Transformer-based Prior-Data Fitted Network, the cardinality of the context set is strictly bounded. Specifically, for TabPFN, its constraint is $(n_{\mathrm{tr}}+n_{\mathrm{test}})p \leq 100,000$ (`https://ux.priorlabs.ai/predict`). In a specific task, the dimension p and n_{test} are always pre-determined. Therefore, the constraint is essentially $n_{\mathrm{tr}} \leq n_{\max}$ for some sample size limit $n_{\max}$.

In summary, transfer learning with TFMs presents two primary challenges: the large volume and the inherent heterogeneity of source data. Because the in-context learning framework constrains the number of training samples during inference, the vast volume of source data must be efficiently summarized into a limited context window. Furthermore, the single-task nature of current TFMs makes them sensitive to internal data heterogeneity. Consequently, it is critical to selectively incorporate only the most informative source data into the transfer learning context.

1.2 Our contributions

To address the challenges of integrating large-scale heterogeneous data into TFMs, we introduce a novel transfer learning framework centered on anchoring and distillation, termed TL-ANDI. Our contributions are summarized as follows.

We propose a three-stage pipeline consisting of anchoring, distillation, and aggregation. The anchoring stage identifies source samples with minimal posterior shift that are highly represen-

tative of the test covariate distribution. The distillation stage summarizes information around these anchor points to effectively reduce predictive uncertainty. Finally, an aggregation step is employed to calibrate the model and mitigate the risk of negative transfer.

Unlike traditional transfer learning approaches that rely on pre-specified functional classes, TL-ANDI leverages the inherent nonlinear modeling capabilities of TFMs. This allows for accurate predictions in complex settings without requiring explicit statistical model specifications. We explicitly incorporate unlabeled test covariate data into the transfer learning process to better align source and target distributions. We establish theoretical guarantees for the proposed distillation mechanism, characterizing its effect on variance reduction.

Furthermore, extensive empirical evaluations across diverse simulation setups and two real-world datasets, spanning both regression and classification tasks, demonstrate that TL-ANDI consistently outperforms established baselines.

1.3 Related works

Transfer learning in conventional statistical models have been extensively studied recently. Recent works have studied transfer learning approaches for nonparametric regression (Cai and Wei, 2021; Reeve et al., 2021) and high-dimensional parametric models (Li et al., 2022; Tian and Feng, 2023; Li et al., 2024) among many others. However, these transfer learning are based on white-box models such that the functional class of the prediction model need to be pre-specified. On the other hand, transfer learning based on black-box models, especially the TFMs in the in-context learning framework, is barely studied.

Tabular Foundation Models have emerged as a powerful alternative, utilizing In-Context Learning to make predictions without task-specific tuning. While early versions like TabPFN (Hollmann et al., 2025) were restricted to small-scale datasets, recent advancements have pushed these boundaries. TabPFN-2.5 (Grinsztajn et al., 2025) has scaled in-context learning capabilities to 50,000 samples and 2000 features, while Qu et al. (2025) introduce a two-stage architecture to handle up to 500,000 samples. Despite these architectural improvements, TFMs can require large computational sources when dealing with large-scale datasets and remain highly sensitive to data heterogeneity and distribution shifts, often leading to performance degradation when source and target distributions diverge significantly. Ye et al. (2025) show that TabPFN v2's performance degrades on both large-scale and high-dimensional datasets but it can be used as feature extractor. Ma et al. (2024) propose a method to combine in-context learning-based retrieval with self supervised learning to train tabular foundation models.

To improve the robustness of TFMs, several adaptation strategies have been proposed. Kang et al. (2025) introduce a framework for tabular data distillation via column-embedding representations to protect privacy. Thomas et al. (2024) propose to retrieve a personalized training dataset as a local context for each test point. However, this method does not work well when posterior shift exists. Helli et al. (2024) address temporal distribution shifts by training a transformer on millions of synthetic datasets where the underlying structural causal models (SCMs) gradually change over time.

1.4 Organization

The remainder of this paper is organized as follows. In Section 2, we introduce the methodological framework of TL-ANDI. Section 3 establishes the theoretical foundations of our approach, providing guarantees for the distillation mechanism. Section 4 presents extensive empirical evaluations, including simulation studies under various distribution shift scenarios. In Section 5, we apply our proposal to two real datasets considering regression and classification tasks, respectively. We conclude with a discussion of our findings and potential avenues for future research in Section 6.

2 Method

We introduce the proposed method in this section. In Section 2.1, we first introduce a residual-based transfer learning framework for TFMs operating under a strict budget of $n_{\max}$ source samples. Subsequently, in Section 2.2, we detail our proposed three-stage methodology to strategically sample and distill the raw source data into this optimal context.

2.1 Transfer Learning using TFMs with one informative source

This section introduces the foundational building block of our methodology: a vanilla transfer learning approach for TFMs. In this initial formulation, we temporarily set aside the complexities associated with large-scale volume and internal source heterogeneity. We simply treat the unified source data as a single domain and model the target domain residuals to calibrate for distribution shifts. Building upon the principles established by Li et al. (2022), we propose a two-stage transfer learning procedure specifically tailored for the black-box operational nature of TFMs.

Algorithm 1: Vanilla Transfer Learning for Black-Box TFMs

Input: Target data $(X^{(\text{tar})}, \boldsymbol{y}^{(\text{tar})})$ of size n_{tar}, source data $(X^{(\text{src})}, \boldsymbol{y}^{(\text{src})})$ of size n_{src}, and test covariates $X^{(\text{test})}$.

Step 1. Learn the source prediction rule:

Prompt the TFM, $\mathcal{P}$, with the source data to predict outcomes for both the target training covariates and the test covariates:

$$\hat{f}^{(\text{src})}(X^{(\text{tar})}, X^{(\text{test})}) = \mathcal{P}\left((X^{(\text{tar})}, X^{(\text{test})}) \mid \mathcal{C}^{(\text{tr})} = \mathcal{D}^{(\text{src})}, \mathcal{C}^{(\text{test})} = \{X^{(\text{tar})}, X^{(\text{test})}\}\right).$$

Step 2. Estimate the target residual bias:

We compute the prediction residuals on the observed target data:

$$\hat{\boldsymbol{z}}^{(\text{tar})} = \boldsymbol{y}^{(\text{tar})} - \hat{f}^{(\text{src})}(X^{(\text{tar})}).$$

Next, we use the TFM to predict the residuals for the test covariates, using the target features and residuals as context:

$$\hat{\boldsymbol{\delta}}(X^{(\text{test})}) = \mathcal{P}\left(X^{(\text{test})} \mid \mathcal{C}^{(tr)} = \{(X^{(\text{tar})}, \hat{\boldsymbol{z}}^{(\text{tar})})\}, \mathcal{C}^{(test)} = \{X^{(\text{test})}\}\right).$$

Final Prediction:

$$\hat{\boldsymbol{y}}(X^{(\text{test})}) = \hat{f}^{(\text{src})}(X^{(\text{test})}) + \hat{\boldsymbol{\delta}}(X^{(\text{test})}).$$

Algorithm 1 operates under the assumption that an informative source dataset is available and adheres to the sample size limit $n_{\text{src}} \leq n_{\max}$. In practice, however, raw source data often significantly exceeds this architectural budget and exhibits internal heterogeneity, which can precipitate negative transfer. To address these dual challenges, the following subsection details our data distillation pipeline, designed to construct a compact and highly informative source context.

2.2 Source Data Distillation

To construct an optimal training context that adheres to the TFM's architectural limits, we propose a source data distillation framework. This framework consists of three integrated stages: (i) identifying representative anchor points within the source domain to align with the target distribution; (ii) applying local label smoothing to reduce predictive variance; and (iii) aggregating predictions across a range of bandwidths via a target-domain calibration process.

2.2.1 Stage 1: Anchor Sampling

The primary objective is to select a subset of $n_{\max}$ "anchor" points, denoted by $\mathring{\boldsymbol{x}}_1, \ldots, \mathring{\boldsymbol{x}}_{n_{\max}}$, from the available source covariates $X^{(\text{src})}$. These anchors are strategically chosen to represent regions of the feature space that are most beneficial for target-task inference. To achieve this, we define sampling weights that account for two distinct types of distribution shifts.

The first type is **covariate shift**, which means that there exist discrepancies in marginal

feature distributions: $p_X^{(\text{src})}(\boldsymbol{x}) \neq p_X^{(\text{tar})}(\boldsymbol{x})$ and (or) $p_X^{(\text{tar})}(\boldsymbol{x}) \neq p_X^{(\text{test})}(\boldsymbol{x})$. Existing works have proposed approaches to adjust for covariate shift based on the idea of density weighting (Shimodaira, 2000; Sugiyama et al., 2008). However, these methods cannot be directly used in TFMs due to their black-box nature. Moreover, existing literature extensively covers covariate shift from source to target, the shift from target to test is rarely addressed.

The second type is **posterior shift**, which means that there exist discrepancies in the conditional distributions: $\mathbb{E}[y^{(\text{src})} \mid \boldsymbol{x}^{(\text{src})} = \boldsymbol{x}] \neq \mathbb{E}[y^{(\text{tar})} \mid \boldsymbol{x}^{(\text{tar})} = \boldsymbol{x}]$. As we reviewed in Section 1.3, existing methods tackle distribution shift in parametric models or in nonparametric classification problems. It is unclear how to detect and deal with posterior shift in black-box models.

To simultaneously detect these two levels of shift, we construct a composite index. We define the covariate shift index $\hat{\alpha}_i^{(\text{CS})}$ to measure the popularity of the i-th source covariate within the test distribution:

$$\hat{\alpha}_i^{(\text{CS})} \propto \frac{\frac{1}{n_{\text{test}}} \sum_{l=1}^{n_{\text{test}}} K_{h_1}(\boldsymbol{x}_l^{(\text{test})} - \boldsymbol{x}_i^{(\text{src})})}{\frac{1}{n_{\text{src}}} \sum_{l=1}^{n_{\text{src}}} K_{h_1}(\boldsymbol{x}_l^{(\text{src})} - \boldsymbol{x}_i^{(\text{src})})}, i = 1, \ldots, n_{\text{src}},$$

where $K_h(\cdot) = h^{-p} K(\cdot/h)$ is a symmetric kernel function and h denotes bandwidth. In practice, one can use Box kernel, Epanechnikov kernel, and so on. This serves as an empirical kernel estimate of the density ratio $p_X^{(\text{test})}(\boldsymbol{x}_i^{(\text{src})})/p_X^{(\text{src})}(\boldsymbol{x}_i^{(\text{src})})$. Higher the $\hat{\alpha}_i^{(\text{CS})}$, more representative the covariate $\boldsymbol{x}_i^{(\text{src})}$ for the test distribution. Note that $\hat{\alpha}_i^{(\text{CS})}$ does not measure the discrepancy between target and source covariate distribution but measures the discrepancy between the test and source covariate distribution. This is because there can be distribution shift between $p_X^{(\text{tar})}(\cdot)$ and $p_X^{(\text{test})}(\cdot)$. In this case, a source sample which is noninformative for target data can be informative for test data.

Second, we define the posterior alignment index $\hat{\alpha}_i^{(\text{PS})}$, to detect regions of minimal posterior shift. We use an estimate of the conditional density ratio $p^{(\text{src})}(y|\boldsymbol{x})/p^{(\text{tar})}(y|\boldsymbol{x})$ to detect posterior shift. Similar ideas have been adopted in Zhao et al. (2026). If the noises are Gaussian, the density ratio has a parametric form

$$\frac{\sigma_{\text{tar}}}{\sigma_{\text{src}}} \exp\left\{-\frac{(y - f^{(\text{src})}(\boldsymbol{x}))^2/\sigma_{\text{src}}^2}{(y - f^{(\text{tar})}(\boldsymbol{x}))^2/\sigma_{\text{tar}}^2}\right\},$$

where σ_{src} and σ_{tar} denote the noise level in the source and target distribution, respectively. In the above expression, $f^{(\text{src})}$ cannot be directly estimated due to the large volume of source data. We propose a simple indicator and tune it through cross-validation. Specifically, by fitting a base target model $\hat{f}^{(\text{tar})}$ (TabPFN trained solely on $\mathcal{D}^{(\text{tar})}$), we evaluate the alignment of source labels with target expectations:

$$\hat{\alpha}_i^{(\text{PS})} \propto \exp\left\{-\frac{(y_i^{(\text{src})} - \hat{f}^{(\text{tar})}(\boldsymbol{x}_i^{(\text{src})}))^2}{2\hat{\sigma}_{\text{tar}}^2}\right\}, \ i = 1, \ldots, n_{\text{src}}$$

where $\hat{\sigma}_{\text{tar}}^2 = \sum_{i=1}^{n_{\text{tar}}} (y_i^{(\text{tar})} - \hat{f}^{(\text{tar})}(\boldsymbol{x}_i^{(\text{tar})}))^2/n_{\text{tar}}$ estimates the noise-level in the target distribution. We introduce a hyper-parameter $\gamma \in [0, 1]$ to balance the effects of the two shifts. The final

sampling weight for the i-th source sample is:

$$\hat{\omega}_i \propto \hat{\alpha}_i^{(\mathrm{CS})} \times (\hat{\alpha}_i^{(\mathrm{PS})})^{\gamma},\ i = 1, \ldots, n_{\mathrm{src}}. \tag{2}$$

As $\hat{\alpha}_i^{(\mathrm{PS})}$ uses the response, the samples selected based on $\hat{\alpha}_i^{(\mathrm{PS})}$ are subject to post-selection effect. That is, the selected sources may have a different model than the original source data. Thanks to the black-box nature of TFMs, our proposal addresses the post-selection bias by performing the calibration step (Step 2 of Algorithm 1).

We denote the selected $n_{\max}$ anchor samples as $\mathring{\boldsymbol{x}}_l$, $l = 1, \ldots, n_{\max}$.

2.2.2 Stage 2: Local Kernel Distillation

Once $\mathring{\boldsymbol{x}}_l$, $l = 1, \ldots, n_{\max}$ are sampled, we consider an aggregated estimate of their corresponding response values to reduce variance. Specifically, we populate their corresponding target values $\mathring{y}_l$ using a Nadaraya-Watson local smoothing operator. For a given bandwidth h_2, the distilled label $\mathring{y}_l$ is a geometrically weighted average of the raw source observations:

$$\mathring{y}_l = \frac{\sum_{i=1}^{n_{\mathrm{src}}} K_{h_2}(\boldsymbol{x}_i^{(\mathrm{src})} - \mathring{\boldsymbol{x}}_l) y_i^{(\mathrm{src})}}{\sum_{i=1}^{n_{\mathrm{src}}} K_{h_2}(\boldsymbol{x}_i^{(\mathrm{src})} - \mathring{\boldsymbol{x}}_l)}. \tag{3}$$

To see why this step can reduce variance, consider a generic non-parametric model $y = f(\boldsymbol{x}) + \epsilon$ with $\mathrm{Var}(\epsilon) = \sigma_{\mathrm{src}}^2$. The variance of the distilled observation $\mathring{y}_l$ is approximately $\sigma_{\mathrm{src}}^2/m_l$, where m_l is the effective number of source points within the local bandwidth h_2. By smoothing out these local fluctuations, we provide the TFM with a "denoised" context that allows it to be more representative of the true predictive function. We will apply Algorithm 1 based on the distilled source data and target data.

2.2.3 Stage 3: Data-Driven Bandwidth Aggregation

Step 1 and Step 2 involve tuning parameters h_1, h_2, and γ by cross-validation. Let $\mathcal{H}$ denote the set of candidate values for bandwidth and Γ denote the set of candidate values for γ.

As detailed in Algorithm 2, we partition the target data $\mathcal{D}^{(\mathrm{tar})}$ into calibration and validation folds. For each choice of tuning parameter, we construct a distilled context $\mathcal{C}^* = \{(\mathring{\boldsymbol{x}}_l, \mathring{y}_l)\}_{l=1}^{n_{\max}}$ and learn a transfer predictor using Algorithm 1. We then select the optimal tuning parameters based on the validation samples.

Algorithm 2: Transfer Learning via ANchoring and DIstillation (TL-ANDI)

Input: Source data $\mathcal{D}^{(\text{src})}$, Target data $\mathcal{D}^{(\text{tar})}$, Test covariates $X^{(\text{test})}$, budget $n_{\max}$, and bandwidth set $\mathcal{H}$.

Output: Final test predictions $\hat{\boldsymbol{y}}^{(\text{test})}$.

Step 0. Data Splitting:

Randomly split $\mathcal{D}^{(\text{tar})}$ into two disjoint subsets: $\mathcal{D}^{(cal)}$ (calibration) and $\mathcal{D}^{(val)}$ (validation), with corresponding index sets $\mathcal{N}^{(cal)}$ and $\mathcal{N}^{(val)}$.

For each $h_1 \in \mathcal{H}, h_2 \in \mathcal{H}, \gamma \in \Gamma$:

Step 1. Anchor Sampling:

Sample covariates $\mathring{\boldsymbol{x}}_1, \ldots, \mathring{\boldsymbol{x}}_{n_{\max}}$ from $X^{(\text{src})}$ without replacement according to the probability weights $\hat{\omega}_i$ defined in (2).

Step 2. Local Kernel Distillation:

- Compute distilled labels $\mathring{y}_l$ for $l = 1, \ldots, n_{\max}$ via Eq. (3) to form the distilled source dataset $\mathcal{D}^{(\text{src})} = \{\mathring{\boldsymbol{x}}_i, \mathring{y}_i\}_{i=1}^{n_{\max}}$.
- Execute Algorithm 1 using $\mathcal{D}^{(\text{src})}$ as the source data and $\mathcal{D}^{(\text{cal})}$ as the target data. Let $\hat{f}(\cdot)$ denote the resulting predictor.
- Generate predictions $\hat{f}_{\boldsymbol{\lambda}}(\boldsymbol{x}_i)$ for all $i \in \mathcal{N}^{(\text{val})}$ and test covariates $X^{(\text{test})}$, where $\boldsymbol{\lambda} = (h_1, h_2, \gamma)^T$ denotes the current tuning parameters.

Step 3. Validation Aggregation:

Select the optimal hyper-parameters $\hat{\boldsymbol{\lambda}}$ that minimize the validation error:

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}:\lambda_1 \in \mathcal{H}, \lambda_2 \in \mathcal{H}, \lambda_3 \in \Gamma}{\arg\min} \sum_{i \in \mathcal{N}^{(val)}} \left(y_i^{(\text{tar})} - \hat{f}_{\boldsymbol{\lambda}}(\boldsymbol{x}_i) \right)^2.$$

Generate the test predictions

$$\hat{y}_i^{(\text{test})} = \hat{f}_{\hat{\boldsymbol{\lambda}}}(\boldsymbol{x}_i^{(\text{test})}), \ i = 1, \ldots, n_{\text{test}}.$$

3 Theoretical guarantees

In this section, we provide a theoretical justification for the proposed methodology, focusing specifically on the efficiency gains achieved through the distillation step. We establish how local kernel smoothing affects the bias-variance tradeoff within the context of transfer learning for TFMs.

Assumption 1. *Assume that the covariates* $\boldsymbol{x}_i^{(\text{src})} \in \mathbb{R}^p$ *are drawn independently from a continuous density* $p_X^{(\text{src})}(\boldsymbol{x})$*, where the support of* $p_X^{(\text{src})}(\boldsymbol{x})$ *is* $[0,1]^d$ *and* $\min_{\boldsymbol{x} \in [0,1]^d} p_X^{(\text{src})}(\boldsymbol{x}) \geq c_0 > 0$*. Assume the noise* $\epsilon_i^{(\text{src})}$ *is independent sub-Gaussian with* $\max_{i \leq n_{\text{src}}} \mathbb{E}[(\epsilon_i^{(\text{src})})^2] \leq \sigma_{\max}^2$*.*

Assumption 2. *Assume that* $f^{(\text{src})}(\cdot)$ *is Hölder smooth such that*

$$\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} |f^{(\text{src})}(\boldsymbol{x}) - f^{(\text{src})}(\boldsymbol{x}')| \leq \theta_1 \|\boldsymbol{x} - \boldsymbol{x}'\|_2^{\theta_2}$$

for some finite θ_1 and some $\theta_2 > 0$. We denote $f^{(\mathrm{src})} \in H(\boldsymbol{\theta})$.

In Assumption 2, we assume $f^{(\mathrm{src})}(\cdot)$ to be Hölder smooth with Hölder norm $\theta_1 > 0$ and smoothness coefficient $\theta_2 > 0$. As $\theta_2 < 1$ is allowed, the source model $f^{(\mathrm{src})}$ is not necessarily smooth.

Assumption 3. *The kernel function $K_h(\cdot)$ satisfies that $K : \mathbb{R}^d \to \mathbb{R}_+$ is bounded, compactly supported, and satisfies*

$$\int_{\mathbb{R}^d} K(u)\, du > 0 \text{ and } \qquad \int_{\mathbb{R}^d} K^2(u)\, du > 0.$$

Assumption 3 puts mild conditions on the kernel function. Suitable choices include box kernels and Epanechnikov kernels.

Lemma 3.1 (Distribution of distilled samples)**.** *Suppose that Assumptions 1, 2, and 3 hold. Then for any given h_2 such that $h_2 = o(1)$ and $nh_2^p \gtrsim \log n$, it holds that*

$$\mathring{y}_l = f^{(\mathrm{src})}(\mathring{\boldsymbol{x}}_l) + \mathring{\Delta}_l + \mathring{\epsilon}_l,$$

where

$$\max_{l \le n_{\max}} |\mathring{\Delta}_l| \le \theta_1 h_2^{\theta_2} \text{ and } \max_{l \le n_{\max}} \mathrm{Var}(\mathring{\epsilon}_l) \le C\sigma_{\max}^2/(n_{\mathrm{src}} h_2^p)$$

for some positive constant C.

We see that the distilled response has variance of order $(n_{\mathrm{src}} h_2^p)^{-1}$. If $n_{\mathrm{src}} h_2^p \to \infty$, then the variance of the distilled response is of smaller order than the variance of the original source response. On the other hand, the distillation step introduces an approximation bias $\mathring{\Delta}_l$, which gets larger as h_2 gets larger. Therefore, we can select the optimal h_2 which balances the bias and variance of the transfer learning estimator. Specially, if $f^{(\mathrm{src})}(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}$ is linear model, then

$$\mathring{y}_l = \bar{\boldsymbol{x}}_l^T \boldsymbol{\beta} + \bar{\epsilon}_l,$$

which still follows a linear model and there is no approximation bias.

4 Simulation Studies

In this section, we benchmark the empirical performance of the proposed TL-ANDI framework against comparable baselines through rigorous simulation studies. We structure our evaluation across the homogeneous source setting in Section 4.1 and the heterogeneous source setting in Section 4.2. Subsequently, Section 4.3 presents ablation studies to validate the individual methodological stages, quantifying the efficacy of anchor sampling and the variance reduction achieved via local kernel distillation respectively.

To establish a rigorous comparison, we consistently evaluate three primary methods: the Target-only baseline, which exclusively utilizes the target data of size n_{tar} for prediction; TL-RAND, a naive transfer learning baseline that handles the oversized source data of size n_{src} by uniformly sampling a subset of budget $n_{\max}$ prior to executing Algorithm 1; and our proposed

TL-ANDI method introduced in Algorithm 2. Across all evaluated methods, we adopt TabPFN-2.5 as the base model. We evaluate the predictive performance by reporting the Mean Squared Error (MSE). Specifically, for test data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n_{\text{test}}}$ and a generic prediction rule $\hat{f}(\cdot)$, we report

$$\text{MSE}(\hat{f}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (f^*(\boldsymbol{x}_i) - \hat{f}(\boldsymbol{x}_i))^2, \tag{4}$$

where $f^*(\boldsymbol{x}_i)$ denotes the true conditional mean function for the target model.

We choose proper tuning parameters for all evaluated methods by cross-validation and report their best respective predictive performances. In practice, the hyper-parameter γ for anchor sampling is selected via a grid search across the candidate set $\gamma \in \{0, 0.01, 0.05, 0.2, 0.5, 0.8, 1\}$. For the computation of the covariate shift index $\hat{\alpha}_i^{(\text{CS})}$, a Box kernel is employed for the symmetric kernel function $K_{h_1}(\cdot)$. Additionally, the Epanechnikov kernel is adopted during data distillation. Finally, for the data-driven bandwidth aggregation phase, the target dataset is partitioned into a calibration subset $\mathcal{N}^{(\text{cal})}$ containing 70% of target samples and a validation subset $\mathcal{N}^{(\text{val})}$ containing 30% of target samples.

We consider a feature dimension of $p = 10$, a target sample size of $n_{\text{tar}} = 150$, a source sample size of $n_{\text{src}} = 20000$, a test sample size of $n_{\text{test}} = 1000$, and budget $n_{\max} \in \{500, 1000\}$.

To evaluate the proposed method under varying degrees of domain discrepancy, we design four specific experimental settings over homogeneous source data and heterogeneous source data.

4.1 Homogeneous Source

In this subsection, we consider the settings with homogeneous source data.

For the source data $\mathcal{D}^{(\text{src})}$, the covariates are *i.i.d.* sampled from $x_{i,j}^{(\text{src})} \sim \mathcal{U}(-1, 1)$. We consider the response model

$$y_i^{(\text{src})} = f^{(\text{src})}(\boldsymbol{x}_i^{(\text{src})}) + \epsilon_i^{(\text{src})},$$

where $\epsilon_i^{(src)} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ are independent noise terms. We define the underlying true function $f^{(\text{src})}(\boldsymbol{x}_i^{(\text{src})})$ as an additive model which contains linear, quadratic, and absolute value terms:

$$f^{(\text{src})}(\boldsymbol{x}) = \sum_{j=1}^{10} \beta_j x_j + \sum_{j=1}^{10} \beta_{j+10} x_j^2 + \sum_{j=1}^{10} \beta_{j+20} v_j(\boldsymbol{x}) + \sum_{j=1}^{5} \beta_{j+30} |x_j|, \tag{5}$$

where β_j, $1 \leq j \leq 35$ are *i.i.d.* drawn from $\mathcal{U}(-1, 1)$ and $(v_1(\boldsymbol{x}), \ldots, v_{10}(\boldsymbol{x})) = \{x_j x_k\}_{1 \leq j < k \leq 5}$ denotes the interaction terms of the first 5 features.

For target data $\mathcal{D}^{(\text{tar})}$, covariates $\boldsymbol{x}_i^{(\text{tar})}$ are drawn from the same distribution as $\boldsymbol{x}_i^{(\text{src})}$. The corresponding response $y_i^{(\text{tar})}$ retains the standard normal additive noise structure, formulated as $y_i^{(\text{tar})} = f^{(\text{tar})}(\boldsymbol{x}_i^{(\text{tar})}) + \epsilon_i^{(\text{tar})}$, where

$$f^{(\text{tar})}(\boldsymbol{x}) = f^{(\text{src})}(\boldsymbol{x}) + \Delta(\boldsymbol{x}).$$

In the following, we evaluate the performance of different methods with different $\Delta(\cdot)$ functions.

(a) **Linear shift.** We first consider linear $\Delta(\boldsymbol{x})$ such that

$$\Delta(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{c},$$

where $c_j \sim \mathcal{U}(-1, 1)$ for $j = 1, \ldots, 5$ and $c_j = 0$ for $j \geq 6$.

(b) **Nonlinear shift.** Next, we consider a nonlinear posterior shift such that

$$\Delta(\boldsymbol{x}) = c_1 x_1 + c_2 x_2^2 + c_3 \sin(2\pi x_3) + c_4 \max(0, x_4) + c_5 (x_1 x_2)$$

where the shift coefficients are sampled uniformly such that $c_j \sim \mathcal{U}(0.2, 0.8)$ for $j = 1, \ldots, 5$.

The test dataset $\mathcal{D}^{(\text{test})}$ is generated following the identical data generating process as the target dataset $\mathcal{D}^{(\text{tar})}$ in settings (a) and (b), respectively.

We evaluate the overall predictive performance of the proposed TL-ANDI method against the Target-only and TL-RAND baselines in settings (a) and (b) and the results are reported in Table 1. The performance of the Target-only method remains invariant to changes in the context budget $n_{\max}$, given that it relies exclusively on the fixed target dataset $\mathcal{D}^{(\text{tar})}$. Both TL-RAND and TL-ANDI report progressively lower MSE as $n_{\max}$ increases, demonstrating their capacity to effectively leverage source data.

In Table 1, the Target-only baseline yields the highest MSE, whereas both TL-RAND and TL-ANDI successfully leverage the source data to achieve better performance. The performance gap between TL-RAND and TL-ANDI is notably narrower under the homogeneous linear shift setting compared to the nonlinear shift setting. This occurs because TL-RAND fails to adapt to complex, nonlinear posterior shifts. Consequently, the proposed TL-ANDI method consistently outperforms both the Target-only and TL-RAND baselines across all homogeneous source settings.

Table 1: Comparison of Mean Squared Error (MSE) among the Target-only baseline, TL-RAND, and our proposed TL-ANDI method under homogeneous source settings. The table presents the mean with standard deviation of the MSE across setting (a) linear and setting (b) nonlinear posterior shifts with context budgets of $n_{\max} = 500$ and $n_{\max} = 1000$. The results are summarized over 50 independent replications.

Setting	$n_{\max}$	**Target-only**	**TL-RAND**	**TL-ANDI**
Linear Shift	500	0.6356 ± 0.1508	0.2987 ± 0.0888	**0.2437 ± 0.0665**
	1000	0.6356 ± 0.1508	0.2227 ± 0.0654	**0.2000 ± 0.0598**
Nonlinear Shift	500	0.7723 ± 0.1785	0.3842 ± 0.0820	**0.3261 ± 0.0702**
	1000	0.7723 ± 0.1785	0.3116 ± 0.0690	**0.2830 ± 0.0639**

4.2 Heterogeneous Source

In this subsection, we consider the settings with heterogeneous source data. Specifically, for the source data $\mathcal{D}^{(\text{src})}$, we consider it to be an integration of two subpopulations.

- For $1 \leq i \leq n_{\text{src}}/2$, the source covariates are drawn from multivariate normal distributions $x_{i,j}^{(\text{src})} \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5^2)$, $1 \leq j \leq 10$. The response $y_i^{(\text{src})}$ are generated via (5) where β_j are *i.i.d.* generated from $\mathcal{U}(-1, 1)$. We denote the corresponding source model by $f^{(\text{src})}(\cdot)$.

- For $n_{\text{src}}/2 < i \leq n_{\text{src}}$, the source covariates are drawn from multivariate normal distributions $x_{i,j}^{(\text{src})} \overset{i.i.d.}{\sim} \mathcal{N}(1, 1)$, $1 \leq j \leq 10$. The response $y_i^{(\text{src})}$ are generated via (5) where β_j are *i.i.d.* generated from $\mathcal{U}(-0.5, 1.5)$. We denote the corresponding source model by $\tilde{f}^{(\text{src})}(\cdot)$.

In the above, we consider a heterogeneous source where there exist both covariate and posterior shifts between the two components.

For the target data, we generate target domain covariates from $x_{i,j}^{(\text{tar})} \sim \mathcal{N}(0, 0.6^2)$. The target model is defined as $f^{(\text{tar})}(\boldsymbol{x}) = f^{(\text{src})}(\boldsymbol{x}) + \Delta(\boldsymbol{x})$, where $\Delta(\boldsymbol{x})$ is defined in settings (a) and (b), and $f^{(\text{src})}(\cdot)$ is the true conditional mean function for the first half of source data. We see that the target data has mild covariate shift and posterior shift from the first half of the source data but has larger distribution shift from the second half of source data. Therefore, the first half of source data are more informative than the second half and can be selected with higher probability. The test dataset $\mathcal{D}^{(\text{test})}$ is generated following the identical data generating process as the target dataset $\mathcal{D}^{(\text{tar})}$ in settings (a) and (b), respectively.

The results are reported in Table 2. TL-RAND performs substantially worse than the Target-only baseline across both $n_{\max} = 500$ and $n_{\max} = 1000$ context budgets, demonstrating severe negative transfer. Expanding the context budget fails to meaningfully alleviate this performance degradation, highlighting the inherent risk of naive random sampling in heterogeneous source settings. In contrast, TL-ANDI consistently achieves the lowest MSE by strategically isolating informative source data from noninformative samples. Ultimately, these results highlight the advantage of TL-ANDI over both relying exclusively on limited target data and employing TL-RAND.

Table 2: Overall performance comparison of Mean Squared Error (MSE) among the Target-only baseline, TL-RAND, and our proposed TL-ANDI method under heterogeneous source settings. The table presents the mean with standard deviation of the MSE across setting (a) linear and setting (b) nonlinear posterior shifts with context budgets of $n_{\max} = 500$ and $n_{\max} = 1000$. The results are summarized over 50 independent replications.

Setting	$n_{\max}$	**Target-only**	**TL-RAND**	**TL-ANDI**
Linear Shift	500	0.9961 ± 0.2426	2.0203 ± 0.7441	**0.6016 ± 0.1428**
	1000	0.9961 ± 0.2426	1.9994 ± 0.7891	**0.5133 ± 0.1286**
Nonlinear Shift	500	1.1393 ± 0.2763	2.0616 ± 0.7191	**0.7239 ± 0.1367**
	1000	1.1393 ± 0.2763	2.0084 ± 0.7738	**0.6029 ± 0.1268**

4.3 Ablation Study

In this subsection, we conduct ablation studies to evaluate the contributions of individual stages under both homogeneous and heterogeneous source settings. To explicitly evaluate the contributions of each stage, we compare the standard random baseline, TL-RAND; the random baseline augmented with our distillation process, TL-RAND (w/ Distill); our proposed anchoring method without distillation, TL-ANDI (w/o Distill); and our complete integrated pipeline, TL-ANDI.

Results for the homogeneous and heterogeneous source settings are presented in Figure 1 and Figure 2, respectively.

We first examine the homogeneous source setting in Figure 1. All evaluated methods demonstrate a reduction in MSE as the context budget $n_{\max}$ increases, regardless of whether local kernel distillation is applied, which is consistent with previous observations. Under the homogeneous linear shift setting, TL-RAND maintains a competitive performance compared with TL-ANDI (w/o Distill), showing that the transfer learning based on naive random sampling can still transfer useful knowledge under mild linear perturbations. The MSE reduction for TL-ANDI in this setting is largely attributable to the data distillation stage, as evidenced by the substantial gap in MSE between TL-ANDI (w/o Distill) and the complete TL-ANDI method. Under the homogeneous nonlinear shift setting, while TL-ANDI (w/o Distill) now maintains a consistent performance advantage over TL-RAND, the profound impact of data distillation remains the same. Again, this distillation step dramatically reduces the MSE for TL-ANDI, ultimately yielding the best overall performance.

Furthermore, we observe that the marginal improvement yielded by distillation is substantially more pronounced at $n_{\max} = 500$ than at $n_{\max} = 1000$ in Figure 1. This is because the limited context leaves the model sensitive to noise under a smaller sample budget. Thus, local kernel distillation leads to a more significant reduction in MSE. Conversely, the TabPFN model internally corrects a portion of the noise through its robust in-context learning capabilities when the budget expands to $n_{\max} = 1000$. As the remaining room for improvement shrinks, the additional variance reduction provided by distillation remains beneficial but becomes less pronounced.

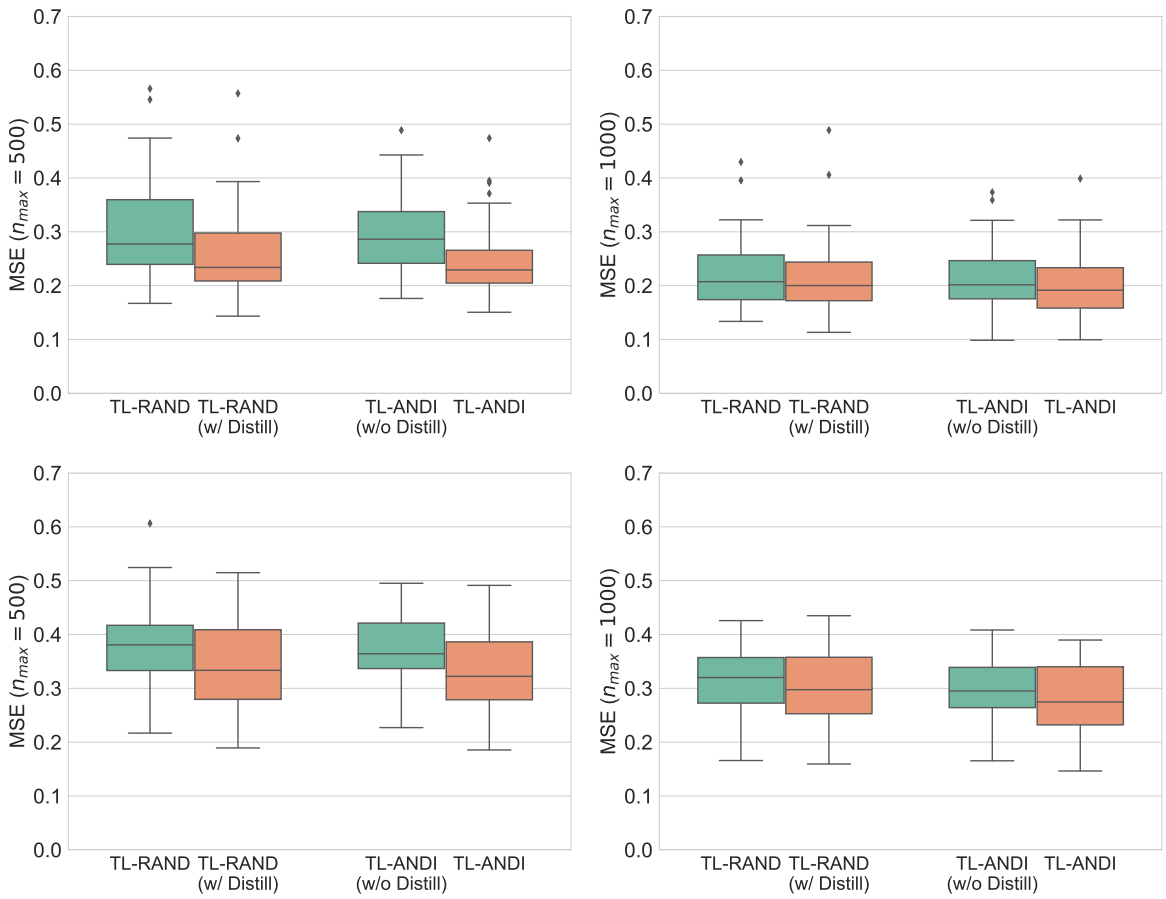

Figure 1: Ablation study under homogeneous source conditions for setting (a) linear shift (first row) and setting (b) nonlinear shift (second row). The boxplots compare Mean Squared Error (MSE) across four configurations: the random baseline TL-RAND, the distilled random baseline TL-RAND (w/ Distill), our proposed anchoring method TL-ANDI (w/o Distill), and the complete pipeline TL-ANDI. The left and right columns correspond to context budgets of $n_{\max} = 500$ and $n_{\max} = 1000$, respectively. Results are summarized over 50 independent replications.

Next, we move on to the heterogeneous source setting in Figure 2. We observe a distinctly different pattern that anchor sampling plays the dominant role in MSE reduction. The massive performance gap between TL-RAND and TL-ANDI (w/o Distill) in the heterogeneous source setting is significantly more pronounced than in the homogeneous source setting. This is because naive TL-RAND suffers from severe negative transfer due to its random sampling mechanism. TL-ANDI (w/o Distill) actively prevents this through anchor sampling, dramatically reducing the MSE. It employs $\hat{\alpha}^{(\mathrm{CS})}$ to manage covariate shifts and leverages target-domain residuals via $\hat{\alpha}^{(\mathrm{PS})}$ for posterior shifts, thereby effectively filtering out non-informative samples. Consequently, this targeted selection avoids the substantial negative transfer that typically plagues TL-RAND. Ultimately, the complete TL-ANDI pipeline further suppresses residual label noise through its distillation mechanism, enabling the combined approach to achieve the best overall predictive performance. Overall, however, the MSE reduction achieved by anchor sampling crucially overwhelms the contribution of data distillation in the heterogeneous source setting.

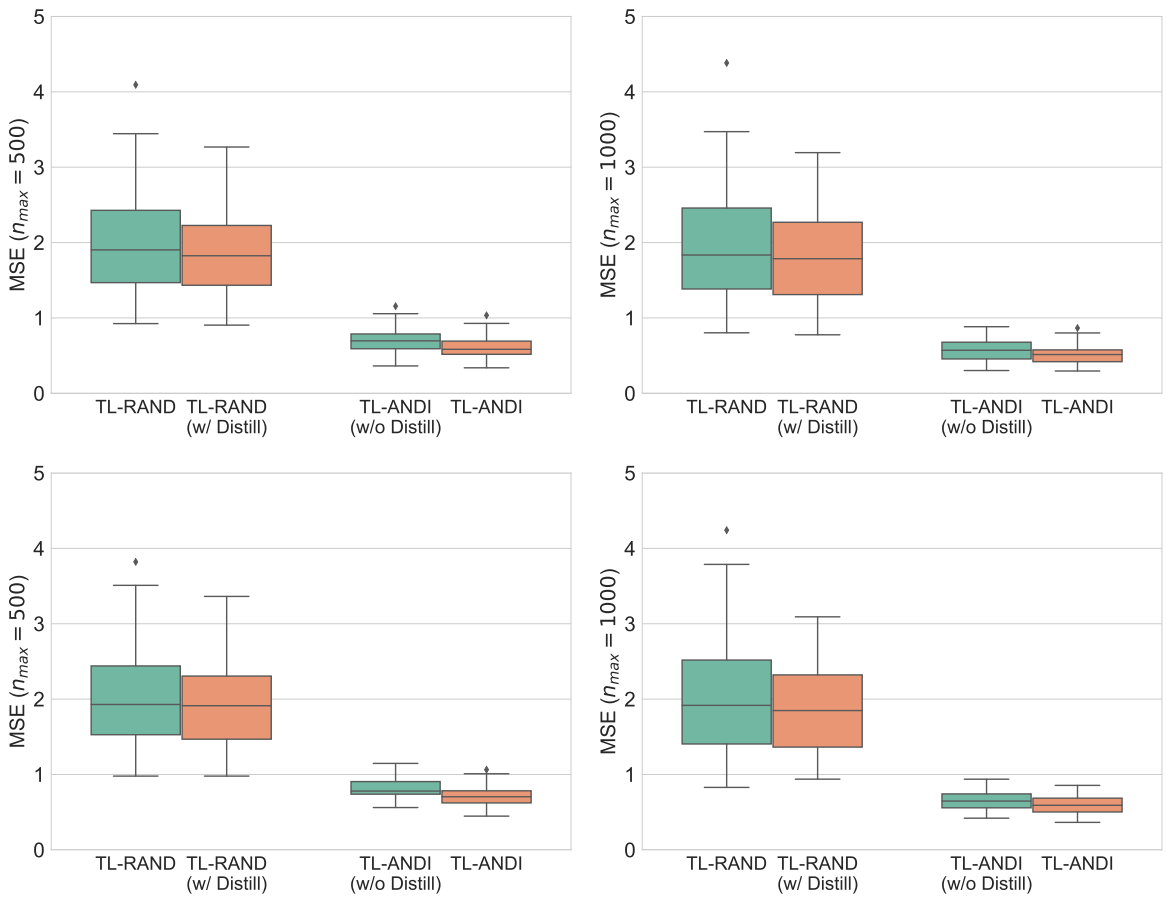

Figure 2: Ablation study under heterogeneous source conditions for setting (a) linear shift (first row) and setting (b) nonlinear shift (second row). The boxplots compare Mean Squared Error (MSE) across four configurations: the random baseline TL-RAND, the distilled random baseline TL-RAND (w/ Distill), our proposed anchoring method TL-ANDI (w/o Distill), and the complete pipeline TL-ANDI. The left and right columns correspond to context budgets of $n_{\max} = 500$ and $n_{\max} = 1000$, respectively. Results are summarized over 50 independent replications.

5 Real Data Experiments

In this section, we evaluate the performance of our proposed transfer learning algorithm in two real datasets considering prediction and classification, respectively.

5.1 Application to California Housing Price Dataset

The California Housing Dataset (https://www.kaggle.com/datasets/camnugent/california-housing-prices/data#) contains aggregate statistics for houses within various California districts based on 1990 census data. The dataset comprises 20,640 observations, with the median house value as the response variable and 9 features. In this analysis, we evaluate and compare the performance of three methods: the Target-only method, TL-RAND, and our proposed TL-ANDI method.

During data pre-processing, we remove all records containing missing values. We also observe that the median house value and housing median age variables are artificially capped at 500,000

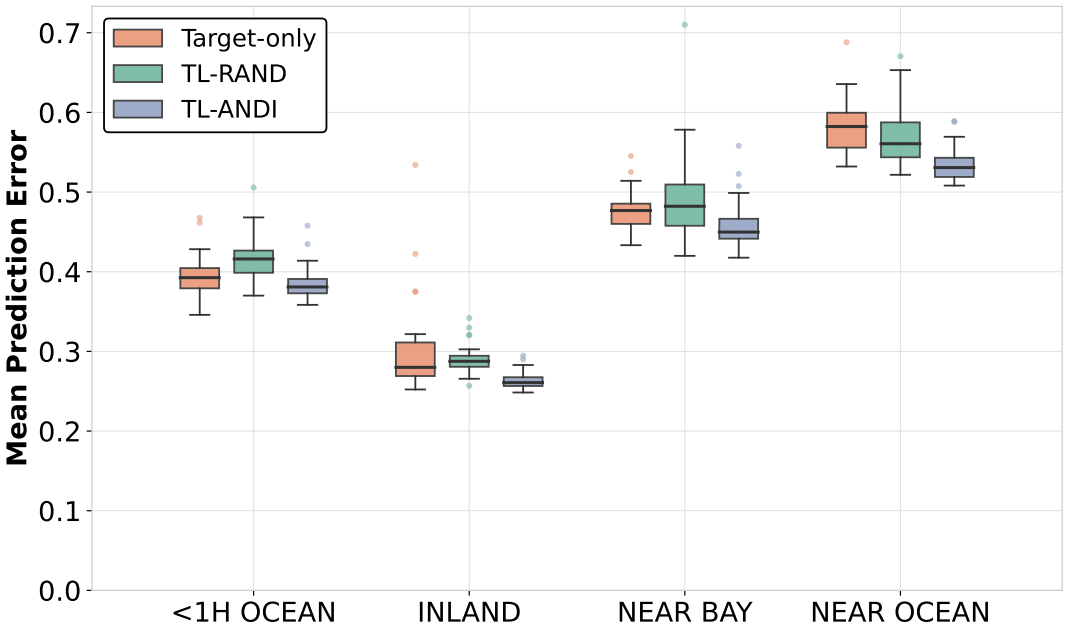

Figure 3: Mean Prediction Error (MPE) of Target-only baseline, TL-RAND, and our proposed TL-ANDI method across four target domains defined by ocean proximity. The results for each method are summarized over 30 independent replications.

and 52, respectively. We exclude these capped observations from our dataset.

To evaluate the robustness and replicability of our method, we categorize the districts based on the ocean proximity feature: "<1H OCEAN", "NEAR BAY", "INLAND", and "NEAR OCEAN". While these categories likely share structural similarities, their underlying predictive models for house values may vary, making them suitable scenarios for transfer learning. We exclude the "ISLAND" category due to its insufficient sample size, whereas the remaining four categories each contain over 2000 observations. We employ a leave-one-out strategy for our experiments; specifically, one category is selected as the target distribution, while the data from the remaining three categories are aggregated to form the source distribution in each iteration. The target sample size is fixed at $n_{\text{tar}} = 100$ with a context budget of $n_{\max} = 1000$, and all remaining samples in the selected category are used for testing. This data-splitting process is repeated 30 times to ensure statistical significance. We define the median house value as the response variable and utilize the remaining 6 features as covariates, excluding longitude and latitude to focus strictly on non-spatial attributes. All features are standardized prior to analysis, and the hyper-parameters are set identically to those used in the simulation studies.

In this experiment, performance is evaluated using the Mean Prediction Error (MPE) on the test data $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n_{\text{test}}}$:

$$\text{MPE}(\hat{f}) = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{f}(\boldsymbol{x}_i))^2, \tag{6}$$

where $\hat{f}(\cdot)$ denotes the prediction rule derived from each method.

Figure 3 demonstrates the Mean Prediction Error (MPE) of the Target-only method, TL-RAND, and the TL-ANDI method for predicting the median house value. We see that the proposed method achieves the smallest prediction error in all the four experiments. When

treating "<1H OCEAN" and "NEAR BAY" categories as the target data, the naive random selection baseline TL-RAND frequently underperforms relative to the Target-only method. This is potentially driven by severe distribution shifts and TL-RAND suffers from negative transfer by blindly pooling heterogeneous source data. Conversely, TL-ANDI consistently maintains competitive performance, indicating the advantage of transfer learning. On the other hand, when treating "INLAND" and "NEAR OCEAN" as target categories, we see that TL-RAND shows smaller prediction errors compared with the Target-only method. Furthermore, TL-ANDI achieves substantially larger performance gains in these two settings. This demonstrates that our proposed method effectively capitalizes on the underlying similarities between the source and target domains, thereby significantly amplifying the positive transfer effect.

Overall, the proposed TL-ANDI method consistently achieves the lowest mean prediction errors across all ocean proximity categories. It yields an average MPE reduction of 6.31% compared to the Target-only method and 7.08% compared to TL-RAND. This empirical evidence validates the practical effectiveness and robustness of our proposed framework in real-world scenarios.

5.2 Application to Diabetes Detection

We consider a classification task in this subsection. The Diabetes Health Indicators Dataset(https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset) contains 70,692 survey responses from the CDC's 2015 Behavioral Risk Factor Surveillance System (BRFSS), a health-related telephone survey. The dataset is perfectly balanced, featuring a 50-50 split between respondents with no diabetes and those with either prediabetes or diabetes. It includes 21 feature variables detailing health-related risk behaviors, chronic health conditions, and the utilization of preventative services. In our analysis, we benchmark our proposed TL-ANDI method against three baselines: the Target-only method, the TL-RAND baseline, and the TabPFN-kNN baseline proposed in Thomas et al. (2024). The TabPFN-kNN baseline operates during the prediction phase by independently retrieving the k-nearest source samples as a unique context for each test sample, relying strictly on covariate distances.

To evaluate our method's performance across heterogeneous data sources, we sequentially treat samples grouped by age and sex as the target domain. Given that the relationship between health risk factors and diabetes prevalence often exhibits heterogeneity across distinct demographic profiles, partitioning the dataset along the age and sex dimensions naturally induces distribution shifts, establishing a suitable transfer learning scenario. The original dataset is split into 13 groups by age. We aggregate them into three broader age clusters: young (groups 1-4), middle-aged (groups 5-8), and old (groups 9-13). By interacting these three age clusters with the binary sex variable, we construct 6 distinct demographic groups (i.e., Young female, Young male, Middle-aged female, Middle-aged male, Old female, Old male). The sample sizes across these six groups range from a minimum of 3315 to a maximum of 21418.

The response variable takes a value of 0 for individuals with no diabetes and 1 for those with prediabetes or diabetes. After defining our target and source domains based on age and sex, we select 10 key features from the remaining 19 variables to serve as covariates. For the Target-only, TL-RAND, and TL-ANDI methods, the raw output provides a continuous estimate of the

probability of having diabetes. We apply a standard classification threshold of 0.5 to determine the final binary prediction. All models are evaluated and compared using mean mis-classification error.

In our experimental setup, one demographic group is isolated as the target and test domain, while all data from the other five categories are aggregated to serve as the source data. We fix the budget $n_{\max} = 1000$, the target size at $n_{\text{tar}} = 100$, and the test sample size at $n_{\text{test}} = 1000$. To ensure the robustness of our evaluation, this random sample splitting process is repeated 30 times. All continuous features are standardized prior to model training, and the hyper-parameters are set identically to those utilized in the simulation studies.

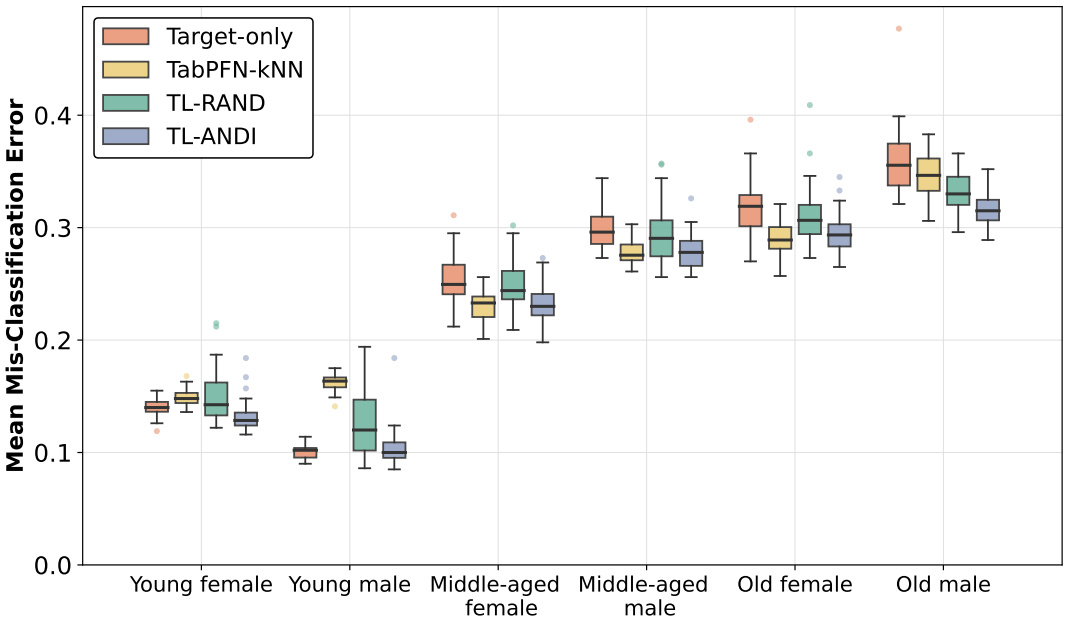

Figure 4: Mean mis-classification error of Target-only baseline, TabPFN-kNN baseline, TL-RAND, and our proposed TL-ANDI method across six demographic target domains in Diabetes Health Indicators Dataset. The results for each method are summarized over 30 independent replications.

Figure 4 reports the average mis-classification errors of the Target-only method, the TabPFN-kNN baseline, the TL-RAND baseline, and our proposed TL-ANDI method across six distinct demographic target domains. Overall, TL-ANDI demonstrates consistently stable and highly competitive or superior performance across all evaluated scenarios. Notably, our method outperforms TL-RAND in every target domain, exhibiting both lower median errors and narrower variance. Our proposal also outperforms TabPFN-kNN in many scenarios, especially in the Young male and Old male groups.

In the "Young female" and "Young male" groups, the Target-only method demonstrates relatively small classification errors in comparison to naive transfer learning methods. This may be due to the existence of a large distribution shift from the young-aged group to older groups so that little information can be transferred. This is because the naive transfer learning problems, such as TabPFN-kNN and TL-RAND perform even worse than the Target-only method. Our proposal remains robust due to its careful selection of anchor points, thereby avoiding negative transfer.

In the middle-aged groups, TabPFN-kNN and TL-ANDI show clear improvement. This potentially indicates that the posterior shift in these groups is mild, meaning that adjusting exclusively for the covariate shift is sufficient for successful knowledge transfer. In the "Old male" group, both TL-RAND and TL-ANDI exhibit significant performance gains through positive transfer. By carefully selecting the most informative source samples, TL-ANDI maximizes this benefit and achieves the lowest overall mean mis-classification error.

In summary, while the Target-only and TabPFN-kNN methods can achieve strong results in specific cases, their success is highly inconsistent. In comparison, our proposed TL-ANDI method consistently delivers robust and superior performance across all demographic scenarios. This again demonstrates the robustness and reliability of our proposed method.

6 Discussion

This work introduces TL-ANDI, a framework designed to overcome the data integration bottlenecks inherent in current Tabular Foundation Models (TFMs). By strategically selecting anchor points to align covariate and posterior distributions—and subsequently distilling these sources into low-variance pseudo-samples—our approach enables TFMs to leverage heterogeneous source data while strictly adhering to architectural context constraints. Theoretical analysis and empirical benchmarks demonstrate that this local distillation strategy provides a robust safeguard against negative transfer, effectively bridging the gap between classical nonparametric smoothing and modern in-context learning.

Despite these advantages, the computational cost of data-driven bandwidth selection remains a hurdle for real-time deployment, suggesting a need for more scalable selection heuristics. Furthermore, as kernel-based methods are sensitive to the curse of dimensionality, future research should explore the integration of dimension reduction or sparse feature selection within the TL-ANDI framework. Developing high-dimensional transfer learning methods that maintain the efficiency of parameter-free, input-space optimization remains a vital direction for expanding the utility of TFMs in complex data environments.

References

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification. *The Annals of Statistics*, 49(1):100–128, 2021.

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.

Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient tabpfn: In-context learning temporal distribution shifts on tabular data. *Advances in Neural Information Processing Systems*, 37:98742–98781, 2024.

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.

Inwon Kang, Parikshit Ram, Yi Zhou, Horst Samulowitz, and Oshani Seneviratne. On learning representations for tabular data distillation. *arXiv preprint arXiv:2501.13905*, 2025.

Sai Li and Linjun Zhang. Personalizing black-box models for nonparametric regression with minimax optimality. *arXiv preprint arXiv:2601.01432*, 2026.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Sai Li, Linjun Zhang, T Tony Cai, and Hongzhe Li. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 119(546):1274–1285, 2024.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse C Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maksims Volkovs. Tabdpt: Scaling tabular foundation models on real data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2024.

Jingang Qu, David HolzmÃžller, GaÃl Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Adaptive transfer learning. *The Annals of Statistics*, 49(6):3618–3649, 2021.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony Caterini. Retrieval & fine-tuning for in-context tabular models. *Advances in Neural Information Processing Systems*, 37:108439–108467, 2024.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global Scientific Publishing, 2010.

Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabpfn v2: Understanding its strengths and extending its capabilities. *arXiv preprint arXiv:2502.17361*, 2025.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

Junlong Zhao, Shengbin Zheng, and Chenlei Leng. Residual importance weighted transfer learning for high-dimensional linear regression. *Journal of the American Statistical Association*, (just-accepted):1–27, 2026.

Appendix

A Proofs

Proof of Lemma 3.1. By (3),

$$\begin{aligned}\mathring{y}_l &= \frac{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l) f^{(\text{src})}(\boldsymbol{x}_i)}{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)} + \frac{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)\epsilon_i}{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)} \\ &= f^{(\text{src})}(\mathring{\boldsymbol{x}}_l) + \underbrace{\frac{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)\{f^{(\text{src})}(\boldsymbol{x}_i) - f^{(\text{src})}(\mathring{\boldsymbol{x}}_l)\}}{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)}}_{T_{l,1}} + \underbrace{\frac{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)\epsilon_i}{\sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)}}_{T_{l,2}}.\end{aligned}$$

As $f^{(\text{src})} \in H(\theta_1, \theta_2)$, we have

$$\max_{1 \le l \le n_{\text{src}}} T_{l,1} \le \theta_1 (\sqrt{p} h_2)^{\theta_2}.$$

On the other hand, conditional on $X^{(\text{src})}$, $T_{l,2}$ is sub-Gaussian with sub-Gaussian norm

$$\text{Var}(T_{l,2} | X^{(\text{src})}) \le \frac{\sigma_{\max}^2}{n_{h_2}(\mathring{\boldsymbol{x}}_l)}$$

for $n_{h_2}(\mathring{\boldsymbol{x}}_l) = \sum_{i=1}^{n_{\text{src}}} K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \mathring{\boldsymbol{x}}_l)$.

To show that

$$\mathbb{P}\left(\inf_{l \le n_{\max}} n_{h_2}(\mathring{\boldsymbol{x}}_l) \ge C n_{\text{src}} h_2^p\right) \ge 1 - \exp\{-c_1 \log n_{\text{src}}\},$$

it suffices to show that for any given $\boldsymbol{x}_i^{(src)}$,

$$\mathbb{P}\left(n_{h_2}(\boldsymbol{x}_i^{(src)}) \ge C n_{\text{src}} h_2^p\right) \ge 1 - \exp\{-c_1 \log n_{\text{src}}\}.$$

Under Assumption 1, for any given $\boldsymbol{x}^*$, we know that

$$\begin{aligned}\mathbb{E}[n_{h_2}(\boldsymbol{x}^*)] &= n_{\text{src}} \mathbb{E}[K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \boldsymbol{x}^*)] \\ &= n_{\text{src}} \int K_{h_2}(\boldsymbol{x}_i^{(\text{src})} - \boldsymbol{x}^*) p_X^{(\text{src})}(\boldsymbol{x}) d\boldsymbol{x}.\end{aligned}$$

A change of variables gives

$$\mathbb{E}[K_h(\boldsymbol{x}_i^{(\mathrm{src})} - \boldsymbol{x}^*)] = \int \frac{1}{h^p} K\left(\frac{\boldsymbol{u} - \boldsymbol{x}^*}{h}\right) p_X^{(\mathrm{src})}(\boldsymbol{u})\, d\boldsymbol{u} = \int K(\boldsymbol{v}) p_X^{(\mathrm{src})}(\boldsymbol{x}^* + h\boldsymbol{v})\, d\boldsymbol{v}.$$

By continuity of $p_X^{(\mathrm{src})}(\cdot)$ at $\boldsymbol{x}^*$,

$$\mathbb{E}[K_h(\boldsymbol{x}_i^{(\mathrm{src})} - \boldsymbol{x}^*)] = p_X^{(\mathrm{src})}(\boldsymbol{x}^*) \int K(\boldsymbol{v})\, d\boldsymbol{v} - O(h) =: c_1 > 0.$$

As $K(\cdot)$ is uniformly bounded, by Bernstein's inequality, it is easy to show that

$$\mathbb{P}\left(h_n(\boldsymbol{x}^*) \geq cnh^p\right) \geq 1 - \exp\{-c_1 \log n\}.$$

As we assume the support of $p_X^{(\mathrm{src})}(\cdot)$ is $[0,1]^d$, we can take a covering of $[0,1]^d$. As $nh_2^p \gtrsim \log n$, it follows from the proof of Lemma C.1 in Li and Zhang (2026), we arrive at desired results. $\square$