BEYOND REMASKING: PROBABILISTIC TABULAR PREDICTION UNDER RECORDING SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

A tabular model can reconstruct hidden entries accurately yet predict the wrong distribution for genuinely missing values. We address recording shift, where selective data collection makes observed outcomes unrepresentative of those requiring prediction. Our approach uses additional information retained in labelled source records and an unlabelled deployment sample, even when that information is unavailable for individual predictions. Building on existing identification results, we develop two probabilistic adaptation routes under conditional stability and source support. Direct reweights learning or labelled contexts. Supported fine-law adaptation (SFLA) combines state-conditional distributions using deployment probabilities. We implement these routes through deployment-targeted flow matching (DT-FM) for joint continuous prediction and through adaptation of frozen tabular foundation models. Our analysis separates correctable population mismatch from information lost at inference and explains how estimation errors affect the benefit of correction. Controlled experiments recover conditional changes invisible to marginal correction and associations lost by point prediction. Experiments on real datasets under imposed recording shifts demonstrate gains over native predictors, with controls isolating the contribution of recording correction.

1 INTRODUCTION

Tabular completion is often trained by masking observed values and reconstructing them, as in masked autoencoders (He et al., 2022), ReMasker (Du et al., 2024), and CSDI (Tashiro et al., 2021). Under selective collection, however, artificially hidden values need not represent genuinely missing ones, even conditional on the same observed inputs. Accurate reconstruction can therefore coexist with a misspecified deployment distribution. We call this discrepancy *recording shift*. In clinical data, for example, a test never ordered and a result still pending may share the same missing-value token, although recording decisions themselves can be predictive (Agniel et al., 2018).

We make this information actionable for probabilistic prediction without observing it for each new case. Labelled source and unlabelled deployment records provide the adaptation signal. *Direct* incorporates recording information through learning weights or context sampling. *Supported fine-law adaptation* (SFLA) mixes state-conditional distributions according to deployment probabilities. We instantiate these routes through deployment-targeted flow matching (DT-FM) for joint continuous prediction and via adaptation of frozen tabular foundation models (TFMs).

Boeken et al. (2023) establish identification using estimation-time information and develop weighting, repeated-regression and doubly robust estimators. Their repeated-regression methods predict means under squared error. For continuous and multivariate outcomes, these means leave intervals, event probabilities and dependence undetermined. We develop full predictive distributions and joint samplers for this setting. SFLA combines fitted distributions directly, preserving their within-state variation and dependence. Changing deployment composition requires no outer outcome regressor. This enables uncertainty assessment and completion of several values with their associations intact.

For TabICLv2, TabDPT and TabPFN (Qu et al., 2026; Ma et al., 2026a; Hollmann et al., 2025; Jäger et al., 2026), the challenge extends beyond encoding a missing input (Figure 1). Pretraining can teach marginalization in the populations it represents. Yet, identical labelled contexts and visible inputs can require opposite predictions in different deployments. The accompanying recording information

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

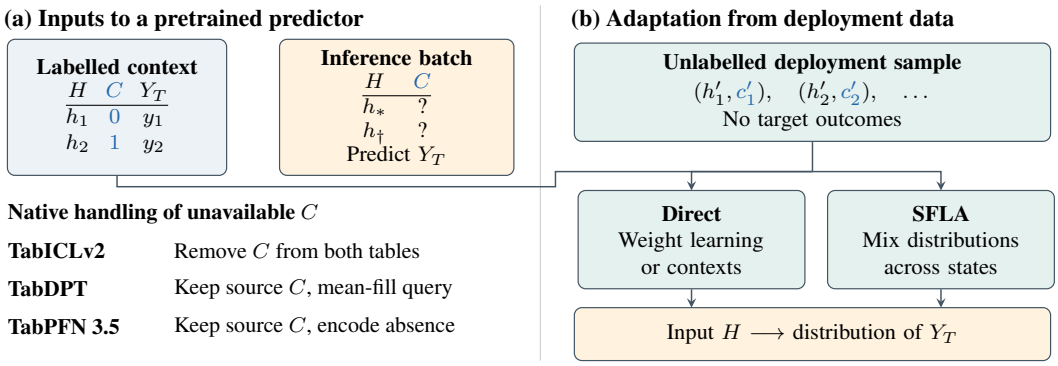


Figure 1: **Recording information guides prediction from available inputs.** H denotes available inputs, C recording information and Y_T targets. (a) The tested APIs handle missing C differently. (b) Direct and SFLA use separate deployment records while predicting from H alone. Interface details: Appendix C.4; clinical example: Appendix A.1.

distinguishes these populations and allows for adaptation with fixed pretrained weights (see Table 2; Appendix B.2).

Expert assessments used in testing or lending may be too costly to obtain for every automated prediction (Boeken et al., 2023), while selective labels can similarly affect group-level assessments (Lakkaraju et al., 2017; Coston et al., 2021; Chang & Wiens, 2024). Related structure appears in trial generalization, where treatment outcomes are observed in the trial but covariates from a target-population sample inform transport beyond the trial, under additional causal and sampling assumptions (Dahabreh et al., 2021; 2023). Table 1 and Appendix G.3 make these designs explicit. When the information is available for every prediction, it can instead enter the model directly.

Building on imputation under MAR (Shannon et al., 2026), selective verification (Begg & Greenes, 1983), auxiliary-variable methods (Collins et al., 2001) and prediction under shift (Shimodaira, 2000; Sugiyama et al., 2007), we make four contributions and evaluate them through controlled experiments and real-data comparisons:

1. **Information limits of native missing-value handling.** We construct deployments with identical labelled contexts, visible inputs and missingness patterns but different optimal predictions. Native prediction cannot distinguish them without additional deployment information (Section 4).
2. **Distributional correction with DT-FM and frozen tabular foundation models.** DT-FM implements deployment-targeted flow training and assembly for joint continuous prediction. Direct and SFLA adapt frozen tabular foundation models through contexts or predictive distributions, without backbone updates. Both retain the same inference inputs (Section 3).
3. **Risk analysis of fitted distributional assembly.** We characterize how component errors make mixture updating help or harm, and separate its gain from the effect of replacing the original predictor with fitted state-conditional components (Section 4).
4. **Controlled and real-data evidence.** Controlled experiments recover conditional changes invisible to marginal correction and associations requiring joint outputs. Real-data comparisons under imposed recording shifts show gains over native predictors, with controls separating correction from context resampling, fitting and model selection (Sections 4–5).

2 Prediction under recording shift

Let H denote the inputs available at inference, Y the outcome to predict, and $C \in \mathcal{C}$ additional information available for estimation but omitted from individual prediction inputs. The outcome may be a vector, and C may contain categorical and continuous variables. Labelled source records contain (H, C, Y). A representative auxiliary sample from deployment contains (H, C) without requiring outcomes. Each individual prediction uses only H.

Table 1: Motivating data designs and concrete examples. Predictions use inputs H; a representative auxiliary sample also contains additional features C. Group evaluation additionally requires membership. These examples illustrate applicability under the assumptions in Appendices A.5 and G.3.

Setting	Concrete example	Labelled source	Auxiliary sample without outcomes
Clinical	Disease risk before a CT scan	Scan results for tested patients	Symptoms and clinical assessments for untested patients
Causal	Blood pressure under a specified treatment	Treatment and follow-up measurements in a randomized trial	Routine measurements and detailed pretreatment assessments in the intended patient population
Fairness	Credit-score error within each applicant group	Repayment outcomes for approved loans	Application details, underwriter assessments and group membership for approved and rejected applicants

Let P_1 and P_0 describe the source and deployment populations, and write $W = (H, C)$. We assume a stable conditional outcome distribution and source support:

$$P_0(dy \mid H, C) = P_1(dy \mid H, C) = Q_F(dy \mid H, C), \qquad P_0^W \ll P_1^W. \tag{1}$$

This *bridge assumption* says that cases with the same (H, C) have the same outcome distribution in both populations; we call Q_F the fine law.. Support requires source labels at the (H, C) values occurring at deployment. Let $\Pi_r(dc \mid h)$ denote the conditional distribution of C given $H = h$ in population r. Prediction from H alone targets

$$Q_{r,H}(dy \mid h) = \int_{\mathcal{C}} Q_F(dy \mid h, c)\, \Pi_r(dc \mid h). \tag{2}$$

The auxiliary samples allow estimation of the deployment distribution Π_0. Changing it can change the prediction even though Q_F remains stable. Our experiments use two states, with probabilities $\pi_r(c \mid h) = \Pi_r(\{c\} \mid h)$. Unlike label shift (Lipton et al., 2018), which preserves inputs given the label, the bridge preserves outcomes given (H, C) (Appendix G.2).

Recorded and artificially hidden outcomes. Let M_j indicate that a value was recorded and S_j that it is revealed to the predictor. Training uses artificial $S \leq M$. $(M_j, S_j) = (1, 0)$ is an observed value withheld as a training target, whereas $(0, 0)$ is genuinely unobserved. A reveal mask conflates these cases, while availability alone may still merge, for example, tests never ordered and results pending. CFMI (Šimkus & Gutmann, 2025) separates recording and training masks and assumes transfer from observed to missing targets. We instead study population differences that can persist after matching masks (Appendix A.2). Auxiliary recording indicators or finer reason codes may enter C, excluding the target's own recording indicator.

For joint completion, a query A specifies observed coordinates $O(A)$ of a response block Y_K and targets Y_T, where $T = K \setminus O(A)$. Then $H = (H^-, A, Y_{O(A)})$, with H^- the available covariates, including missing-value tokens. We use Y_T for the outcome in the following sections. For whole-block recorded-to-missing prediction, $R_K = \prod_{j \in K} M_j$ defines $P_r = P(\cdot \mid R_K = r)$. The bridge becomes $Y_K \perp R_K \mid (H, C)$, allowing recording to remain informative given H. Our real-data experiments impose recording so that hidden outcomes remain available for evaluation (Appendix A).

3 Direct Adaptation and Distributional Assembly

Both routes estimate deployment predictions from source labels and auxiliary records without outcomes. Validation source labels select the final predictor.

Direct adaptation. For deployment e, let $\Gamma_e = dP_e^W / dP_1^W$ be its density ratio relative to the source on $W = (H, C)$. Under the bridge and support assumptions, a fixed predictor satisfies

$$\mathcal{R}_e(Q) = \mathbb{E}_e \ell\{Q(\cdot \mid H), Y_T\} = \mathbb{E}_1[\Gamma_e(H, C) \ell\{Q(\cdot \mid H), Y_T\}]. \tag{3}$$

A source-versus-deployment classifier on (H, C) estimates the ratio, adjusting odds for sample sizes. We fix this estimate during fitting and validation. Weighted loss trains a full model or a head on frozen outputs: Brier loss for class probabilities, CRPS for scalar CDFs (Appendix C.2).

Adaptation of frozen tabular foundation models. For a frozen in-context predictor, *Context Direct* instead samples source examples proportionally to $\widehat{\Gamma}_e$ and averages predictions from L contexts of m examples. Uniform resampling uses the same L, m. Context sampling changes the demonstrations, so its effect is evaluated separately from weighted learning. The latter can also use uniform contexts since the risk identity averages over the same context-generation rule in both populations (Appendix B.10).

Distributional assembly (SFLA). We can fit $\widehat{Q}_F(\cdot \mid h, c)$ from source labels, using a model conditioned on C or separate models for finite states. Alternatively, a frozen in-context model receives labelled examples containing C and a specified state c for prediction. Auxiliary deployment data can be used to estimate $\Pi_e(\cdot \mid h)$. SFLA then forms $\widehat{Q}_{e,H} = \int \widehat{Q}_F(\cdot \mid h, c)\, \widehat{\Pi}_e(dc \mid h)$, combining probabilities or CDFs as in pattern-mixture models and mixtures of experts (Little, 1993; Jacobs et al., 1991). Underrepresented states can thus receive more weight in the deployment prediction. To sample jointly, draw one state from $\widehat{\Pi}_e$ and then the whole target vector from that component. *Source assembly* uses identical components with $\widehat{\Pi}_1$. Its comparison with SFLA measures the effect of changing the mixture probabilities. Appendix C specifies the estimators and model interfaces.

Deployment-targeted flow matching. With Gaussian Z_0, $t \sim \mathrm{Unif}(0,1)$ and $Z_t = (1-t)Z_0 + tY_T$, flow matching (Lipman et al., 2023) and rectified flow (Liu et al., 2023) regress a velocity field on $U = Y_T - Z_0$. DT-FM implements the two routes through

$$\mathcal{L}_0(\theta) = \mathbb{E}_1[\Gamma_0(H, C)\|v_\theta(Z_t, t, H) - U\|^2], \tag{4}$$

$$\mathcal{L}_F(\theta) = \mathbb{E}_1[\|v_\theta(Z_t, t, H, C) - U\|^2]. \tag{5}$$

Direct minimizes $\mathcal{L}_0$ and integrates the field given H. SFLA fits with $\mathcal{L}_F$, draws $C \sim \widehat{\Pi}_0(\cdot \mid h)$, and integrates the field given (H, C). Unit source-to-deployment weights recover source training. Appendix B.8 relates these objectives to native CFMI fitting.

Choosing the prediction. Each candidate Q_k comprises its fitted model, contexts, and sampler. We score it on separate source validation rows using

$$\widehat{R}_0(k) = |\mathcal{V}|^{-1} \sum_{i \in \mathcal{V}} \widehat{\Gamma}_0(h_i, c_i)\widehat{\ell}\{Q_k(\cdot \mid h_i), y_{T,i}\}. \tag{6}$$

The score is CRPS for scalar predictions and energy for joint samples. Our flow rule returns the earliest eligible checkpoint whose paired difference from the best is within one standard error. Appendix C.6 specifies the candidate libraries and comparison with held-out flow loss. Shared Direct reuses a model across deployment mixtures (Appendix C.2).

4 Analysis of Distributional Adaptation

Recording information may identify a deployment distribution without revealing each individual's state. We distinguish this information limit from errors in fitted predictors, and then examine changes in dependence. We use vector Brier loss for classification and the energy score for joint prediction (Gneiting & Raftery, 2007). For independent $X, X' \sim Q$ with finite first moments, $\mathrm{ES}(Q, y) = \mathbb{E}\|X - y\| - \frac{1}{2}\mathbb{E}\|X - X'\|$; its scalar form is CRPS. Their excess risk $D_\ell(P, Q) = \mathbb{E}_{Y \sim P}[\ell(Q, Y) - \ell(P, Y)]$ equals $\|\mu(P) - \mu(Q)\|^2$, where μ averages under mixtures (Sejdinovic et al., 2013). Appendix B.1 gives the representations, normalizations and proofs.

4.1 Which prediction errors can adaptation reduce??

For inputs V, the *Bayes risk* $\mathcal{R}_0^\star(V) = \mathbb{E}_0\ell\{P_0(\cdot \mid V), Y_T\}$ is the minimum expected loss among predictions using V. Here $V = H$ or (H, C).

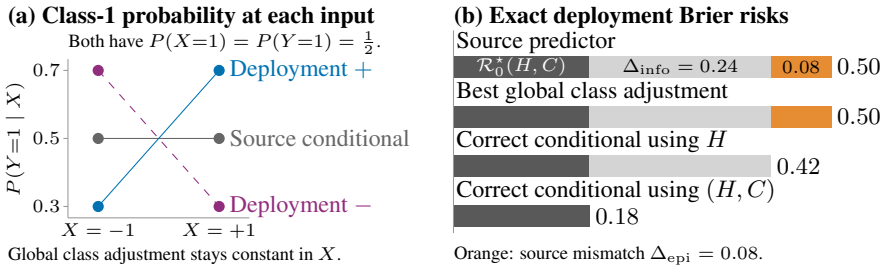
Proposition 1 (Conditional projection)**.** *Under* (1) *and square-integrable embeddings, a forecast fitted independently of the evaluation draw and using only H satisfies*

$$\mathbb{E}_0 D_\ell(Q_F, \widehat{Q}_H) = \underbrace{\mathbb{E}_0 D_\ell(Q_F, Q_{0,H})}_{\Delta^\ell_{\mathrm{info}}} + \mathbb{E}_0 D_\ell(Q_{0,H}, \widehat{Q}_H). \tag{7}$$

The first term is the irreducible loss from omitting C at inference time, while correction then acts on the second term. For the source predictor $Q_{1,H}$, this second term is the population mismatch $\Delta^{\ell}_{\text{epi}}$. With two states at a fixed query, probabilities p_r and separation $d = D_\ell(Q^{(0)}, Q^{(1)})$, the terms are $\Delta^{\ell}_{\text{epi}} = (p_0 - p_1)^2 d$ and $\Delta^{\ell}_{\text{info}} = p_0(1-p_0)d$. Thus, $\mathcal{R}^\star_0(H) = \mathcal{R}^\star_0(H, C) + \Delta^{\ell}_{\text{info}}$.

Identical marginals, different conditionals. Table 2 compares two deployments with identical source labels and marginals $P(H)$, $P(C)$ and $P(Y)$, but opposite associations between H and C. With $H = (X, Z)$, equiprobable $X = \pm 1$ and independent noise Z, their class probabilities are $P_0(Y = 1 \mid H) = 1/2 \pm 0.2X$. Despite having different optimal predictions, the two deployments are indistinguishable from the labelled contexts and test-time inputs, including all mask variables. Consequently, any predictor restricted to these inputs has average vector Brier risk at least 0.50 across the two deployments, while the deployment-specific H-only Bayes predictors each attain risk 0.42. This lower bound is compatible with a predictor being optimal on either deployment individually and applies equally to learned and frozen models (Appendix B.2). Auxiliary (H, C) records distinguish the deployments and enable a correction.

Table 2: Conditional correction under unchanged marginals. All methods predict without the individual's C. Top: exact class probabilities and attainable risks. Below: mean vector Brier risk over both deployments and sixteen source samples. Predictive/Unrelated describes whether C predicts Y given H. Original uses H only; Native uses the released interface. Bold marks lowest displayed means per model and condition. Repeated regression uses the logistic components and all inference features. Context Direct and Uniform use 205 examples per context; Original, Native and components use 512. For paired comparisons and global-correction/CRUMB controls, see Appendix D.1.



	TabICLv2		TabDPT		TabPFN 3.5	
Procedure	Predictive	Unrelated	Predictive	Unrelated	Predictive	Unrelated
Original (H only)	0.5032	**0.5025**	0.5015	0.5010	0.5015	0.5011
Native	0.5032	0.5026	0.5081	**0.5005**	0.5005	**0.5009**
Uniform contexts	0.5320	0.5338	0.5094	0.5109	0.5058	0.5067
Context Direct	0.4581	0.5332	0.4388	0.5155	0.4342	0.5079
Source assembly	0.5017	0.5032	0.5018	0.5010	0.5018	0.5011
SFLA	**0.4212**	0.5036	**0.4214**	0.5011	**0.4214**	0.5013
Attainable risk using H	.4200	.5000	.4200	.5000	.4200	.5000

Comparator	Predictive	Unrelated
DistPFN (TabICLv2)	0.5035	0.5027
DistPFN-T (TabICLv2)	0.5025	0.5020
Logistic: original	0.5031	0.5027
Logistic: weighted Direct	0.4228	0.5034
Logistic: SFLA	0.4218	0.5030
Logistic: repeated regression	0.4220	0.5030

SFLA approaches the attainable risk for all three TFMs, while Context Direct improves each native predictor. DistPFN's prior and temperature adjustments (Lee, 2026) operate on batches that are indistinguishable across deployments and a global class multiplier cannot represent the opposite corrections required at $X = -1$ and $X = 1$ (Appendix B.3). Weighted logistic and repeated regression approach the optimum as well. A binary mean specifies a Bernoulli distribution, making

repeated regression a probabilistic comparator here. Continuous means alone leave uncertainty and dependence unspecified (Appendix D.1.2). In the unrelated-state control, correction increases risk, most strongly for Context Direct and Uniform context.

4.2 When does mixture correction help?

Changing the fitted predictor can help or hurt before recording probabilities are updated. Let B be the original, S source assembly and F SFLA. On the same deployment cases,

$$\mathcal{R}_0(B) - \mathcal{R}_0(F) = \underbrace{\mathcal{R}_0(B) - \mathcal{R}_0(S)}_{\text{build source mixture}} + \underbrace{\mathcal{R}_0(S) - \mathcal{R}_0(F)}_{\text{update mixture weights}}. \tag{8}$$

The first comparison replaces the original with state-conditional predictions mixed using source probabilities. The second changes only those probabilities. At a fixed query, write $\mu_c, \widehat{\mu}_c$ for true and fitted component representations, $\delta = \mu_1 - \mu_0$ and $\widehat{\delta} = \widehat{\mu}_1 - \widehat{\mu}_0$. At source probability p, the mixture error is $\varepsilon = (1-p)(\widehat{\mu}_0 - \mu_0) + p(\widehat{\mu}_1 - \mu_1)$.

Proposition 2 (Effect of a mixture update)**.** *Let $G(x)$ be source-assembly risk minus updated-assembly risk, both evaluated at supported target probability $p + x$. Updating the same fitted components from exact probability p to $p + x$ gives*

$$G(x) = -2x\langle \varepsilon, \widehat{\delta} \rangle + x^2\{\|\delta\|^2 - \|\widehat{\delta} - \delta\|^2\}. \tag{9}$$

Exact components give gain $x^2 d$, while fitting error adds a linear term, so opposite shifts can help or harm, as the fixed-library experiments show (Figure 2a,b). With fixed components and exact source probabilities, estimating deployment probabilities adds a variance penalty $v\|\widehat{\delta}\|^2$ (Appendix B.7). For a smooth TabICLv2 classification task, source labels and unlabelled deployment records predict gain 0.166, compared with 0.173 observed at target prior 0.2 (panel c). Global prior adjustment captures most of this gain because the marginal class distribution changes (Appendix E.2). With unchanged unconditional marginals and estimated probabilities, a separate continuous experiment finds mixture-update gains of $0.0313/0.0314$ normalized energy for ordinary FM/matched CFMI, with positive paired intervals, Gaussian controls achieve similar risks (Appendix D.2).

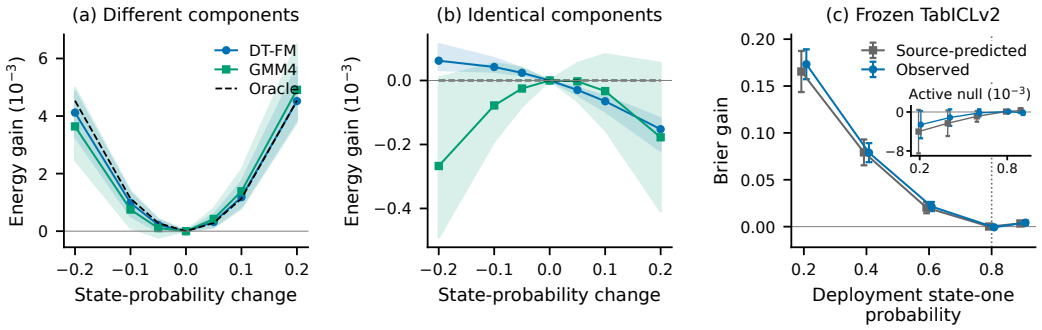


Figure 2: Fitted-error response to mixture updating. Each fit reuses a fixed component library across compositions. (a,b) Energy gains with exact mixture probabilities: ranges over three fits, on separate scales. (c) Predicted and observed Brier gains with estimated probabilities: pointwise 95% intervals over sixteen procedures. The inset uses unrelated recording states; the dotted line marks no shift.

4.3 Why predict a joint distribution?

Suppose we have four responses whose marginal distributions are standard Gaussian in both recording states, but opposite within-pair product moments. Recording then favors negative dependence in source labels and positive dependence among missing outcomes. Mixture updating largely removes the association bias for joint TabICLv2, DT-FM, DiffPuter (Zhang et al., 2025) and GMM4 (Figure 3). TabICLv2 draws one shared state and samples coordinates sequentially, conditioning each on that state, the inputs and earlier draws (Appendix D.3).

Squared-error-optimal point predictions are zero and remove all variation. Independently recombining sampled coordinates preserves their empirical marginals but removes dependence, so both lose the association. Product moments expose this failure when energy is less sensitive to it (Pinson & Tastu, 2013). Indeed, CFMI's energy-selected checkpoint has greater product bias than its final checkpoint.

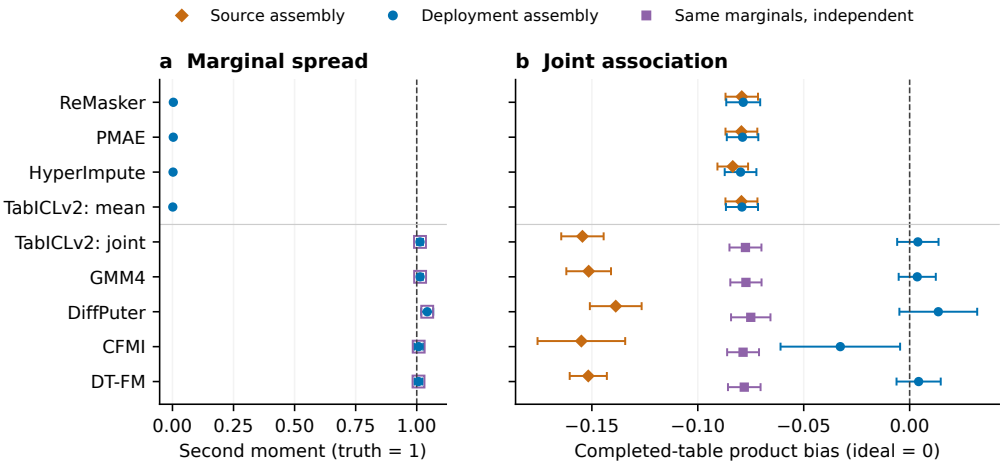


Figure 3: Recovering the association with joint predictions. Known source and deployment probabilities isolate mixture correction on fixed components; individual C is unavailable to the assembled predictor. Squares independently recombine those samples to isolate dependence. Point predictions appear above the divider. Pair (Y_1, Y_2), with pointwise 95% intervals over sixteen procedures. Both pairs and Direct comparisons are reported in Appendix D.3.

5 Experiments

We evaluate source-to-deployment correction, learning and selection on real datasets, followed by joint completion without shift. Deployment outcomes are used only for evaluation.

5.1 Adaptation across applications

Table 3 compares fixed procedures under five imposed recording designs. Beijing, Household and Metro omit station, meter activity and calendar information from inference, whereas Gas and Appliances use response-derived states whose outcome-free availability is assumed experimentally. Because recording depends only on state, the bridge holds by consruction.

Original and Native assess competitiveness. Native TabDPT and TabPFN retain C in labelled examples and use their default missing-feature handling. Source assembly versus SFLA isolates mixture updating, while Uniform versus Context Direct isolates context targeting. CRUMB (Heredge et al., 2026) uses our 205-example-context reimplementation. Flows use 2,816 source labels and the other families 2,560 (Appendix F.1). For flows, both predictive-score and flow-loss validation improve checkpoint selection, with neither criterion uniformly better. Additional fits reproduce these validation gains on the same records. All flow results use source-selected checkpoints (Appendix F.3).

Correction improves native prediction. SFLA improves on native TabPFN 3.5 by .0130 CRPS on Household and .0081 on Metro, with positive paired intervals. TabDPT likewise improves on Native on Household. TabICLv2 SFLA also outperforms both CRUMB variants on Household. Native TabPFN remains stronger on Beijing and Appliances (Appendix F.1).

Mixture updating explains the Metro gains. Updating fixed components reduces CRPS for every family on Metro and Household. On Metro, it also improves every original predictor. TabICLv2 and TabDPT exceed the prespecified .005 useful-gain threshold. Household's gains concentrate on

Table 3: Recording correction across model families: mean CRPS ↓. Updating fixed components lowers mean CRPS for all six displayed families on Household and Metro. Below: pretrained-model update gains with paired, pointwise 95% temporal-group intervals conditional on fits. Bold denotes column minima, and underlying within-family minima. All models omit individual C. †TabICLv2 Original and Native both use H only. Protocols, total gains and controls are in Appendix F.1.

		State defined by response interactions		State defined by other recorded variables		
Model	Procedure	Gas Turbine	Appliances	Beijing	Household	Metro
TabICLv2	Native†	0.2008	0.1748	0.4655	0.3793	0.2098
	Uniform contexts	0.1961	**0.1700**	0.4413	0.3885	0.2231
	Context Direct	0.2044	0.1791	**0.4321**	0.3818	0.2169
	Source assembly	0.1981	0.1728	0.4590	0.3833	0.2123
	SFLA	0.1984	0.1721	0.4587	0.3675	0.1865
	CRUMB	0.2008	0.1825	0.4600	0.3931	0.2112
TabDPT	Original (H only)	0.1908	0.1836	0.4789	0.3808	0.2084
	Native	0.1975	0.1818	0.4575	0.3955	0.1898
	Uniform contexts	0.1883	0.1811	0.4402	0.3820	0.2183
	Context Direct	0.1930	0.1898	0.4353	0.3854	0.2159
	Source assembly	0.1911	0.1809	0.4679	0.3836	0.2113
	SFLA	0.1911	0.1802	0.4668	0.3667	**0.1857**
TabPFN 3.5	Original (H only)	0.1831	0.1750	0.4673	0.3814	0.2108
	Native	0.1822	0.1724	0.4423	0.3785	0.1955
	Uniform contexts	**0.1800**	0.1731	0.4546	0.3786	0.2125
	Context Direct	0.1865	0.1835	0.4479	0.3677	0.1965
	Source assembly	0.1819	0.1759	0.4531	0.3827	0.2130
	SFLA	0.1825	0.1756	0.4525	**0.3655**	0.1874
Quantile trees	Original	0.2199	0.2344	0.4522	0.4068	0.2141
	Weighted Direct	0.2249	0.2273	0.4579	0.3926	0.2006
	Source assembly	0.2195	0.2345	0.4530	0.4082	0.2142
	SFLA	0.2197	0.2348	0.4513	0.3899	0.1895
DT-FM	Original	0.2281	0.2481	0.4637	0.3933	0.2265
	Weighted Direct	0.2476	0.2367	0.4953	0.3965	0.2123
	Source assembly	0.2215	0.2456	0.4839	0.4019	0.2262
	SFLA	0.2255	0.2402	0.4801	0.3964	0.1988
CFMI	Original	0.2314	0.2217	0.4652	0.4541	0.2171
	Weighted Direct	0.2287	0.2234	0.4660	0.4050	0.2165
	Source assembly	0.2201	0.2250	0.4863	0.4277	0.2139
	SFLA	0.2207	0.2219	0.4862	0.4049	0.2021

CRPS gain ↑: source assembly − SFLA; mean [95% interval]

Model	Gas Turbine	Appliances	Beijing	Household	Metro
TabICLv2	−.0003 [−.0013, .0008]	.0007 [.0002, .0012]	.0003 [−.0033, .0038]	.0158 [.0082, .0234]	.0258 [.0174, .0341]
TabDPT	−.00036‡ [−.00220, .00147]	.00054‡ [−.00002, .00109]	.0011 [−.0033, .0055]	.0169 [.0092, .0246]	.0256 [.0175, .0338]
TabPFN 3.5	−.0007 [−.0019, .0006]	.0003 [−.0005, .0012]	.0005 [−.0053, .0064]	.0172 [.0096, .0249]	.0256 [.0172, .0341]

active power, while DT-FM's update recovers only partly its loss from fitting state-specific predictors (Appendix F.1.2).

The controls distinguish other sources of improvement. On Beijing, source assembly and uniform contexts explain much of the improvement, while equal-rate controls leave almost no mixture-update benefit. Gas and Appliances have mixed marginal and joint results. For mean prediction, repeated regression remains competitive. With the same fitted trees, assembly has an RMSE advantage only on Household (Appendix F.1.1).

5.2 Correction in the selected predictor

On Metro, replacing source assembly with SFLA in a fixed library improves the selected prediction by $0.0256/0.0235$ CRPS for TabICLv2/TabDPT, with positive paired intervals and both outperform the equal-label originals. Because the weighted head remains available in both libraries, this isolates the added mixture correction. On Beijing, selection returns uniform contexts whilst Household mostly retains the same head (Appendix F.2.2).

For fixed prediction candidates, recording weights can instead alter which predictor is chosen at validation. Among five TabPFN 3.5 candidates including Native, weighted validation improves over ordinary validation by $0.0134/0.0155$ CRPS on Household/Metro at equal label budgets. Household's selected predictor also beats full-context Native by 0.0085 $[0.0024, 0.0146]$. Native-relative gains remain unresolved on Metro and Beijing; Beijing PM2.5 loses under equal rates (Appendix F.2.2). Reserving validation labels reduces candidate context sizes. On independent Electricity data, the four-candidate TabICLv2/TabDPT procedure finds no clear gain over equal-label originals and misses its useful-gain threshold (Appendix F.2.2).

Weighting both training and stopping improves TabICLv2 Household active power by 0.0271 CRPS $[0.0148, 0.0393]$ with initialization, contexts, labels and training budget fixed. Changing either alone leaves the initialization selected. Weighted heads also improve active power over source-loss heads, but worsen reactive power. For further comparisons, see Appendix F.2.

5.3 Joint completion without recording shift

Without recording shift ($\Gamma_0 = 1$), selected flows have the lowest mean energy in seven of eight queries. Adaptive MIWAE (Mattei & Frellsen, 2019) leads California's four-coordinate query. Table 4 shows two-coordinate results; Appendix F.4 reports all eight.

Table 4: Joint completion: normalized energy, mean (SD across three fits). Bold marks lowest displayed means. Validation source labels select checkpoints. For details, see Appendix F.4.

		Energy/$\sqrt{2}\downarrow$: mean (fit SD)			
Model	Checkpoint / sampling rule	California housing	Protein structure	SGEMM performance	Tetouan power
DT-FM	Final checkpoint	**0.2497** (0.0162)	**0.1154** (0.0010)	**0.00246** (0.00006)	0.4667 (0.0314)
DT-FM	Score-selected	**0.2497** (0.0162)	**0.1154** (0.0010)	**0.00246** (0.00006)	0.3870 (0.0285)
CFMI	Original (100 steps)	0.2580 (0.0137)	0.1186 (0.0009)	0.00251 (0.00005)	0.4863 (0.0394)
CFMI	Selected (400 steps)	0.2580 (0.0136)	0.1186 (0.0008)	**0.00246** (0.00008)	**0.3838** (0.0375)
DiffPuter	Original (training loss)	0.2758 (0.0201)	0.1326 (0.0049)	0.01115 (0.00106)	0.4283 (0.0231)
DiffPuter	Score-selected	0.2821 (0.0240)	0.1324 (0.0046)	0.01063 (0.00143)	0.4313 (0.0298)
Copula	Fixed recipe	0.3371 (0.0127)	0.1974 (0.0008)	0.07379 (0.00133)	0.3968 (0.0226)
GMM8	Without support filter	0.2821 (0.0159)	0.1349 (0.0016)	0.00423 (0.00010)	0.4747 (0.0581)
Stochastic MICE	Fixed recipe	0.3331 (0.0179)	0.1840 (0.0044)	0.00276 (0.00012)	0.3993 (0.0070)
MIWAE	Adaptive SIR	0.2523 (0.0146)	0.1187 (0.0022)	0.00301 (0.00017)	0.4199 (0.0117)

With selectively recorded responses, flows also have lower covariance error than ReMasker, PMAE and HyperImpute on Protein and Tetouan. ReMasker leads on California (Appendix Table F.22).

6 Discussion and limitations

Direct and SFLA use recording information to adapt predictive distributions without changing inference inputs, with controls separating correction from resampling, component fitting and validation. Identification requires the bridge, support and representative auxiliary records. Context Direct and SFLA retain gains under tested bridge departures, though their ranking can reverse (Appendix E.4). Natural missingness and independent gains from a selected procedure remain to be established, whilst possible pretraining exposure limits claims of evaluation of entirely unseen datasets.

REPRODUCIBILITY STATEMENT

Appendix A specifies the prediction populations and information available to each estimator. Appendix B proves the score identities; Appendix C gives the algorithms, selection rules and scoring conventions. Appendices D–F describe data processing, fitting budgets, uncertainty units and the reported experiments, including unsuccessful corrections. Fitting and prediction costs are reported in Appendix C.8.

ETHICS STATEMENT

The PhysioNet experiment uses retrospective clinical measurements under an imposed recording mechanism. It does not evaluate a clinical intervention or establish performance under naturally occurring missingness.

AI USE STATEMENT

Generative AI tools assisted with research-method and experiment-design feedback, development and checking of mathematical arguments and proofs, implementation and debugging, development of synthetic data-generating mechanisms, result interpretation, literature search, and drafting and editing text, figures, tables and LaTeX. This included OpenAI Codex and Claude Code.

REFERENCES

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ*, 361:k1479, 2018. doi: 10.1136/bmj.k1479. URL https://www.bmj.com/content/361/bmj.k1479.

Ibrahim Alabdulmohsin, Nicole Chiou, Alexander D'Amour, Arthur Gretton, Sanmi Koyejo, Matt J. Kusner, Stephen R. Pfohl, Olawale Salaudeen, Jessica Schrouff, and Katherine Tsai. Adapting to latent subgroup shifts via concepts and proxies. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pp. 9637–9661, 2023. URL https://proceedings.mlr.press/v206/alabdulmohsin23a.html.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/e3d3db07c1bfb63e1d0b998996de1d12-Paper-Conference.pdf.

Colin B. Begg and Robert A. Greenes. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, 39(1):207–215, 1983. doi: 10.2307/2530820. URL https://pubmed.ncbi.nlm.nih.gov/6871349/.

P. Boeken, Noud de Kroon, Mathijs de Jong, Joris M. Mooij, and Onno Zoeter. Correcting for selection bias and missing response in regression using privileged information. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 195–205, 2023. URL https://proceedings.mlr.press/v216/boeken23a.html.

Trenton Chang and Jenna Wiens. From biased selective labels to pseudo-labels: An expectation-maximization framework for learning from biased decisions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6286–6324, 2024. URL https://proceedings.mlr.press/v235/chang24e.html.

Jian Chen, Zhehao Li, and Xiaojie Mao. Learning with selectively labeled data from multiple decision-makers. *arXiv preprint arXiv:2306.07566*, 2025. URL https://arxiv.org/abs/2306.07566v4. Version 4, revised 27 May 2025; first posted in 2023.

Linda M. Collins, Joseph L. Schafer, and Chi-Ming Kam. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4):330–351, 2001. doi: 10.1037/1082-989X.6.4.330.

Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2144–2155, 2021. URL https://proceedings.mlr.press/v139/coston21a.html.

Issa J. Dahabreh, Sebastien J-P. A. Haneuse, James M. Robins, Sarah E. Robertson, Ashley L. Buchanan, Elizabeth A. Stuart, and Miguel A. Hernán. Study designs for extending causal inferences from a randomized trial to a target population. *American Journal of Epidemiology*, 190 (8):1632–1642, 2021. doi: 10.1093/aje/kwaa270. URL https://doi.org/10.1093/aje/kwaa270.

Issa J. Dahabreh, Sarah E. Robertson, Lucia C. Petito, Miguel A. Hernán, and Jon A. Steingrimsson. Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, 79(2):1057–1072, 2023. doi: 10.1111/biom.13716. URL https://doi.org/10.1111/biom.13716.

Hugh Dance, Johnny Xi, Peter Orbanz, and Benjamin Bloem-Reddy. Debiased counterfactual generation via flow matching from observations. *arXiv preprint arXiv:2605.07665*, 2026. URL https://arxiv.org/abs/2605.07665.

Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KI9NqjLVDT.

Moonjung Eo, Min-Kook Suh, Hye-Seung Cho, Jiwon Kim, Seoyoon Kim, Sangjun Nam, and Soonyoung Lee. Exaone tabular 1.0: Technical report. 2026.

Zhengyu Fang, Zhimeng Jiang, Huiyuan Chen, Xiao Li, and Jing Li. Understanding and mitigating memorization in diffusion models for tabular data. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 2025. URL https://arxiv.org/abs/2412.11044.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *arXiv preprint arXiv:2505.16959*, 2025. URL https://arxiv.org/abs/2505.16959.

Jake Fawkes, Nic Fishman, Mel Andrews, and Zachary C. Lipton. The fragility of fairness: Causal sensitivity analysis for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f7be3ebca4980b59fe3f665011115395-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270. URL https://doi.org/10.1002/qj.2270.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/people/raftery/Research/PDF/Gneiting2007jasa.pdf.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2111.06377.

Jamie Heredge, Mattia J. Villani, Pranav Deshpande, Akshay Seshadri, and Niraj Kumar. CRUMB: Efficient prior fitted network inference via distributionally matched context batching. *arXiv preprint arXiv:2606.11473*, 2026. URL https://arxiv.org/abs/2606.11473.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL https://miguelhernan.org/whatifbook.

John Hogue. Metro interstate traffic volume, 2019. URL https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume.

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*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://www.nature.com/articles/s41586-024-08328-6.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-MIWAE: Deep generative modelling with missing not at random data. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.12871.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79. URL https://www.cs.toronto.edu/~hinton/absps/jjnh91.pdf.

Benjamin Jäger, Nick Erickson, Léo Grinsztajn, Felix Birkel, Klemens Flöge, Oscar Key, Kürşat Kaya, Jonas Kübler, Adèle Frankel, Tobias Schröder, et al. TabPFN-3.5: Technical report. *arXiv preprint arXiv:2609.17895*, 2026. URL https://arxiv.org/abs/2609.17895.

Daniel Jarrett, Bogdan C. Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. HyperImpute: Generalized iterative imputation with automatic model selection. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9916–9937, 2022. URL https://proceedings.mlr.press/v162/jarrett22a.html.

Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.02557.

Nathan Kallus and Angela Zhou. Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2439–2448, 2018. URL https://proceedings.mlr.press/v80/kallus18a.html.

Fairoz Nower Khan, Nabuat Zaman Nahim, and Peizhong Ju. Flow matching with missing data. *arXiv preprint arXiv:2607.28698*, 2026. URL https://arxiv.org/abs/2607.28698.

Jae Kwang Kim, Seho Park, Yilin Chen, and Changbao Wu. Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3):941–963, 2021. doi: 10.1111/rssa.12696. URL https://doi.org/10.1111/rssa.12696.

Jungkyu Kim, Kibok Lee, and Taeyoung Park. To predict or not to predict? proportionally masked autoencoders for tabular data imputation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017. doi: 10.1145/3097983.3098066. URL https://cs.stanford.edu/~jure/pubs/contraction-kdd17.pdf.

Seunghan Lee. Mitigating label shift in tabular in-context learning via test-time posterior adjustment. *arXiv preprint arXiv:2605.04363*, 2026.

LimiX Team. LimiX: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025. Version 2, 7 November 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. doi: 10.1080/01621459.1993.10594302.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Alex Luedtke and Kenji Fukumizu. DoubleGen: Debiased generative modeling of counterfactuals. *arXiv preprint arXiv:2509.16842*, 2025. URL https://arxiv.org/abs/2509.16842.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Hamidreza Kamkari, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. *arXiv preprint arXiv:2410.18164*, 2026a. URL https://arxiv.org/abs/2410.18164v3. Version 3, January 17; first posted in 2024.

Yuchen Ma, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel. DiffPO: A causal diffusion model for learning distributions of potential outcomes. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2410.08924.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7f8cabaf2de70e1a9d3eb187f02bd58c-Paper-Conference.pdf.

Pierre-Alexandre Mattei and Jes Frellsen. MIWAE: Deep generative modelling and imputation of incomplete data sets. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4413–4423, 2019. URL https://proceedings.mlr.press/v97/mattei19a.html.

Valentyn Melnychuk and Stefan Feuerriegel. GDR-learners: Orthogonal learning of generative models for potential outcomes. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.22953.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Pierre Pinson and Julija Tastu. Discrimination ability of the energy score. DTU Compute Technical Report 2013-15, Technical University of Denmark, 2013. URL https://orbit.dtu.dk/en/publications/discrimination-ability-of-the-energy-score/.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026. URL https://arxiv.org/abs/2602.11139.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ff997469ac66cf893c4183efeb22212a-Paper-Conference.pdf.

Sarah E. Robertson, Jon A. Steingrimsson, Nina R. Joyce, Elizabeth A. Stuart, and Issa J. Dahabreh. Estimating subgroup effects in generalizability and transportability analyses. *American Journal of Epidemiology*, 193(1):149–158, 2024. doi: 10.1093/aje/kwac036. URL https://doi.org/10.1093/aje/kwac036.

Patrick Rockenschaub, Zhicong Xian, Alireza Zamanian, Marta Piperno, Octavia-Andreea Ciora, Elisabeth Pachl, and Narges Ahmidi. Robust prediction under missingness shifts. *arXiv preprint arXiv:2406.16484*, 2024.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581. URL https://doi.org/10.1093/biomet/63.3.581.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

Luke Shannon, Song Liu, and Katarzyna Reluga. Distribution shift in missing data imputation: A risk-based perspective and importance-weighted correction under MAR. *arXiv preprint arXiv:2602.06713*, 2026. doi: 10.48550/arXiv.2602.06713. URL https://arxiv.org/abs/2602.06713.

Xinwei Shen and Nicolai Meinshausen. Engression: Extrapolation through the lens of distributional regression. *arXiv preprint arXiv:2307.00835*, 2023. URL https://arxiv.org/abs/2307.00835.

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. doi: 10.1016/S0378-3758(00)00115-4.

Vaidotas Šimkus and Michael U. Gutmann. CFMI: Flow matching for missing data imputation. *arXiv preprint arXiv:2506.09258*, 2025. URL https://arxiv.org/abs/2506.09258.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007. URL https://jmlr.csail.mit.edu/papers/v8/sugiyama07a.html.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony Caterini. Retrieval and fine-tuning for in-context tabular models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.05207.

Lawson Ung, Tyler J. VanderWeele, and Issa J. Dahabreh. Generalizing and transporting causal inferences from randomized trials in the presence of trial engagement effects. *Epidemiology*, 36(4): 500–510, 2025. doi: 10.1097/EDE.0000000000001863. URL https://doi.org/10.1097/EDE.0000000000001863.

Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

Jiahao Yu, Qizhen Ying, Leyang Wang, Ziyue Jiang, and Song Liu. Missing data imputation by reducing mutual information with rectified flows. *arXiv preprint arXiv:2505.11749*, 2025. URL https://arxiv.org/abs/2505.11749.

Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip S. Yu. DiffPuter: Empowering diffusion models for missing data imputation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9e2ad37b8899366767df9319efc6f4de-Abstract-Conference.html.

Bo Zheng, Yudong Chen, Zihua Xiong, Shuai Fang, Peidong He, Yang Yang, and Sheng Guo. Masktab: Scalable masked tabular pretraining with scaling laws and distillation for industrial classification. *arXiv preprint arXiv:2605.11408*, 2026. URL https://arxiv.org/abs/2605.11408.

CONTENTS

A Prediction targets and information assumptions 16
B Proofs and score identities 21
C Estimation and evaluation 30
D Information and joint prediction 39
E Fitted recording correction 47
F Prediction on real datasets 50
G Related work and scope 71

READING THE SUPPLEMENT

Appendices A–B develop the assumptions and theory. Appendix C specifies estimation and scoring. The experiments follow the main argument: D tests the information needed for correction and joint prediction, E examines estimation error, and F compares performance on real datasets. Appendix G relates the methods to prior work, develops the causal interpretation and states the evidential limits

Scores and comparisons. Lower risk is better. A gain subtracts the adapted risk from its named reference, so positive values favor adaptation. For SFLA, original minus source assembly measures fitting state-specific predictions, source assembly minus SFLA measures updating their probabilities, and original minus SFLA is the total gain. Risk bolding identifies minima within the comparison named by the table; displayed ties are included where specified. Main-table underlines identify the best displayed mean within a model family. When gains are bold, their intervals exclude zero; negative bold gains denote losses.

Uncertainty and controls. Unless stated otherwise, intervals are paired, pointwise 95% intervals. CRPS and energy contrasts use paired scores; RMSE contrasts use the stated delta method on paired group summaries. Intervals across independently fitted procedures also include fitting variation. Fit SD is the standard deviation across fits; a range reports their minimum and maximum. Queries, draws and compositions sharing a fit do not add independent repetitions.

Throughout, source data determine the predictions and deployment outcomes evaluate them. Appendix G.4 gives the scope of these conclusions.

A PREDICTION TARGETS AND INFORMATION ASSUMPTIONS

Recording correction links labelled source cases to the target prediction population. Its required information, assumptions and weights depend on how those populations are defined. We make these choices explicit below and illustrate them with clinical and fairness examples. Recording shift is our main application. The same ideas apply more generally when conditioning on (H, C) gives a stable outcome distribution across populations with overlapping support.

A.1 PREDICTION INPUTS AND TARGETS

Section 2 defines the inference input $H = (H^-, A, Y_{O(A)})$, the target $Y_{T(A)}$ and the additional recording information C. Classification has no observed target coordinates, $O(A) = \varnothing$, and its labels can be selective even when all features are observed. Remasking reveals only recorded coordinates ($S \leq M$). Whether an explicit mask adds information depends on what the values, tokens and query already reveal, so comparisons concern the complete input.

Figure A.1 shows records with the same H and different C. In both, the non-target test X_2 is unavailable and has the same availability indicator, while C distinguishes "not requested" from "pending". Target outcomes and their recording indicators are excluded from the inputs. Prediction from H therefore averages over the unknown status and assigns both records the same distribution. A comparator given C measures the additional value of observing the individual's status (Proposition 1).

A.2 WHAT REMASKING LEARNS

An imputer is intended to predict missing values. The distribution identified by its objective depends on which values can provide labels. Fix a query and include all visible values and mask information

	Covariates X		Response block Y_K		
	Age X_1	Other test X_2	Y_1	Y_2	Y_3
Stored row, $C = 0$	58	Not requested	140	?	?
Stored row, $C = 1$	58	Result pending	140	?	?
After fixed preprocessing	58	NA	140	?	?

Input map: both reasons for the other test's absence become NA.
Query A**:** reveal Y_1; predict (Y_2, Y_3) together.
$K = \{1, 2, 3\}$, $O(A) = \{1\}$, $T(A) = \{2, 3\}$.
$H^- = (58, \mathrm{NA})$, $H = (H^-, A, 140)$, $W = (H, C)$.

Figure A.1: **Recording information lost in preprocessing.** The covariates are $X = (X_1, X_2)$, where X_2 is a non-target test unavailable in both records. Its value is unavailable in both records. The state C records why the test value is unavailable, but preprocessing maps both records to the same input H. Source and auxiliary deployment records retain C. One response is observed and two are predicted jointly. Question marks denote values unavailable at inference time, even if later available as fitting labels.

in H. If P_1 denotes the eligible labelled population after query sampling, then an unrestricted unweighted strictly proper-score objective has optimum

$$q_1^*(\cdot \mid H) := \arg\min_q \mathbb{E}_1 \ell\{q(\cdot \mid H), Y_T\} = P_1(\cdot \mid H) = Q_{1,H} \quad \text{a.s.} \tag{A.1}$$

With squared-error reconstruction the optimum is the conditional mean $\mathbb{E}_1[Y_T \mid H]$. Conditioning the loss on H proves both statements. Flow and diffusion objectives learn distributions of eligible endpoints through surrogate losses. We distinguish these fitting objectives from proper scores of the final sampler.

ReMasker (Du et al., 2024) reconstructs artificially hidden observed entries. CSDI (Tashiro et al., 2021) also selects supervised targets from observed values. Matching artificial masks to deployment can match the query while leaving this population difference unchanged. With incomplete source data, each query has its own label-eligibility event and corresponding ratio (Appendix C.2).

CFMI's mask construction. CFMI (Šimkus & Gutmann, 2025) explicitly separates original availability M from conditioning and target masks s_c, s_t:

$$s_c + s_t = M, \qquad s_c \odot s_t = 0.$$

In our notation, training uses $s_c = S$ and $s_t = M - S$. For $M = (1, 1, 0)$, a possible split is $s_c = (1, 0, 0)$ and $s_t = (0, 1, 0)$. The second coordinate supplies a supervised target. The third is originally missing and enters neither set. CFMI evolves noise on target coordinates and averages squared velocity error over those coordinates. At imputation, it sets $s_c = M$ and $s_t = 1 - M$ to generate the genuinely missing values.

The released vector-field wrapper includes s_c as an input by default and optionally includes s_t. The target mask also restricts the evolving coordinates and output.[1] During training, the pair (s_c, s_t) determines M. The information retained by the complete input therefore matters more than the number of explicit mask channels. CFMI already distinguishes the two processes operationally and acknowledges the assumption needed to generalize from observed training targets to missing outcomes.

For a fixed, supported conditioning/target pattern, CFMI's splitting-induced distribution in its Equation (8) supplies the endpoints for the field loss in Equation (9). Its population conditional is therefore that of the eligible training examples. The target-count normalization balances task losses. Without labelled support for an inference pattern, the objective does not identify its conditional law; population correction additionally requires a source-to-deployment ratio. A target's own recording indicator does not solve this problem, since its unrecorded stratum has no supervised labels. We instead use supported auxiliary states C, such as recording information on other variables, whose role in population correction is distinct from the masks defining eligible training targets.

[1] https://github.com/vsimkus/cfmi, `imp_cfm/models/cfm.py` and `neural_nets.py`.

Correcting the population. The desired conditional is $Q_{0,H}$ for the cases needing predictions. Under the bridge, the populations share Q_F but can differ in their state probabilities:

$$Q_{0,H} - Q_{1,H} = \sum_c \{\pi_0(c \mid H) - \pi_1(c \mid H)\} Q_F(\cdot \mid H, c).$$

Even an exact source conditional can consequently predict the wrong population at the same input H. Weighted Direct estimates the full population ratio and changes the learning objective. SFLA estimates deployment state probabilities and changes the mixture. Both use recording information that can be unavailable for an individual inference case. Table 2 constructs two deployments for which source labels and deployment marginals alone cannot identify the required correction. Flow models (such as CFMI) can be instantiated to implement either route: recording-weighted field learning gives a Direct procedure, while supported C-conditional flows provide SFLA components. This assembly requires predictions conditional on the auxiliary state, in addition to the masks defining each completion task. Appendix B.8 derives the weighted population field. The matched flow comparisons hold the selection and solver policies fixed within each family, and reuse the same components for both mixtures (Appendix F.3). They test correction within the same flow family. The information bound in Appendix B.2 separately establishes what identical source and inference inputs cannot identify, regardless of model capacity. Native imputers can recover the required conditional when their observed-to-missing transfer assumption holds; our bridge supplies a route when stability requires the additional state C.

Other imputation objectives use different identifying assumptions. MIWAE (Mattei & Frellsen, 2019) fits an observed-data likelihood under ignorable missingness. The extension not-MIWAE (Ipsen et al., 2021) adds a missingness model. Importance-weighted imputation (Shannon et al., 2026) explicitly corrects population risk. Retaining both masks can preserve information, while transport still requires an assumption relating labelled and prediction populations.

Two empirical questions. Artificial withholding, also called amputation, can create an evaluation task or define a selectively recorded population. CFMI, for example, evaluates imposed MCAR and MAR missingness. Our distinction concerns which population receives predictions and how it relates to the available labels: unrecorded cases and all eligible cases generally require different ratios (Appendix A.4). When $Q_{1,H} = Q_{0,H}$, population correction is unnecessary. The complete-source benchmark tests estimation of this common conditional (Appendix F.4). Recording-shift studies test the additional benefit of changing the population, with source assembly and matched unweighted predictors separating correction from construction and selection (Appendices D–F).

A.3 The bridge assumption

Let P_1 be the labelled source population and P_0 the deployment population, including its query policy. The bridge in (1) requires a common outcome law given (H, C). The support assumption $P_0^{H,C} \ll P_1^{H,C}$ requires the source to cover the deployment inputs.

The populations may be defined directly, or through a recording event. For whole-block recorded-to-missing prediction, set $R_K = \prod_{j \in K} M_j$, $P_1 = P(\cdot \mid R_K = 1)$ and $P_0 = P(\cdot \mid R_K = 0)$. Source cases have every response label, whereas deployment cases have at least one missing coordinate. The bridge is $Y_K \perp R_K \mid (H, C)$, which may fail after C is omitted. Here R_K defines the populations and is distinct from C. Revealing Y_O changes a partial query's target without changing a study's $R_K = 1$ source. Admitting any row with $Y_{T(A)}$ recorded instead defines a different source and requires a corresponding bridge and ratio (Appendix C.2).

For a single response with all conditioning variables observed, the bridge is the usual MAR condition (Rubin, 1976). In a partly observed table, it applies to the specified pattern and target. Outcome-dependent recording that remains after conditioning on (H, C) violates this bridge and requires additional identifying assumptions.

For partial queries, the bridge is required for the actual inference inputs, including the acquisition pattern and revealed values. Stability under a different information set does not establish this condition (Appendix C.2). Delayed outcomes can later become training labels, changing the source and the required correction. They remain excluded from the inputs available at the historical prediction time.

A.4 Population weights

Let $P_\star$ describe eligible cases before recording, $\rho(h,c) = P_\star(R_K = 1 \mid h, c)$ and $p = P_\star(R_K = 1)$. Bayes' rule gives the normalized ratios on (H, C):

Labelled source	Population receiving predictions	Ratio
Recorded cases	Unrecorded cases	$\frac{p}{1-p}\frac{1-\rho}{\rho}$
Recorded cases	All eligible cases	p/ρ
All eligible historical cases	Cases pending at the prediction time	$(1-\rho)/(1-p)$
All eligible cases	The same eligible population	1

Each ratio has source expectation one. The third row requires historical labels for early and late cases, a common eligible-input population, and a history-to-pending bridge. Discarding late labels instead gives the first row. The second row covers selectively labelled prediction under $Y_T \perp R_K \mid (H, C)$ and positive ρ on the target support, as in our imposed Adult recording task. Predicting outcomes under a different decision additionally needs a potential-outcome target and its identifying assumptions (Appendix G).

To derive the ratios, Bayes' rule gives

$$P_1^W(dw) = \frac{\rho(w)}{p} P_\star^W(dw), \qquad P_0^W(dw) = \frac{1-\rho(w)}{1-p} P_\star^W(dw).$$

Dividing these measures gives the first ratio. Dividing $P_\star^W$ by P_1^W gives the second. Taking $P_\star^W$ as source gives the third. They are density ratios of probability measures, proving normalization where the stated support condition holds.

Alternatively, we can combine representative outcome-free source and deployment samples. If D indicates deployment membership in that pooled sample, $\eta = P(D = 1)$ is its sampling fraction and $q(h,c) = P(D = 1 \mid h, c)$, then

$$\Gamma(h,c) = \frac{1-\eta}{\eta} \frac{q(h,c)}{1-q(h,c)}. \tag{A.2}$$

Bayes' rule gives $q(w)/(1-q(w)) = \{\eta/(1-\eta)\}\{dP_0^W/dP_1^W\}(w)$. The sampling fraction η is set by the sampling design and can differ from the population recording rate. For finite states, the full ratio factorizes as $\Gamma(h,c) = \{dP_0^H/dP_1^H\}(h)\pi_0(c \mid h)/\pi_1(c \mid h)$. The state ratio alone omits changes in the distribution of H. For continuous or mixed states, the corresponding factor is the conditional Radon–Nikodym derivative $d\Pi_0(\cdot \mid h)/d\Pi_1(\cdot \mid h)$, evaluated at c. Source-versus-deployment classification estimates the full ratio directly in each case.

A.5 Auxiliary information and applications

SFLA estimates the conditional state distribution $\Pi_0(\cdot \mid h)$ from representative deployment (H, C) observations, without Y_T. A representative audit can collect recording information for a subset of deployment cases, or a separate data process can join that information to the same inference features. Retrospective records must represent the relevant deployment population. The sample needs both H and C to estimate their conditional relationship. If C is available for an individual query, prediction from (H, C) is instead the richer-input alternative. For finite states, write $\pi_0(c \mid h) = \Pi_0(\{c\} \mid h)$; pooled counts suffice when these probabilities are constant across h. Observations of H alone need not identify the mixture. The unchanged-marginals experiment (Section 4.1) isolates this distinction.

General recording information. Take the spaces of H, C and Y_T to be standard Borel, which includes finite categories and vectors of categorical and real-valued variables. Regular conditional distributions then exist. Under the bridge and source support, the tower property gives

$$Q_{r,H}(dy_T \mid h) = \int Q_F(dy_T \mid h, c)\, \Pi_r(dc \mid h). \tag{A.3}$$

Conditioning first on (H, C) proves the identity; integrating the probability kernel preserves normalization. Estimation and sampling are specified in Appendix C.1.

Selective testing: disease risk before a CT scan. Consider predicting a disease that a CT scan can establish. Let H contain age and routine vital signs, C the clinician's additional symptom assessment, Y the disease indicator, and R whether the scan was performed. Source records contain (H, C, Y) for tested patients ($R = 1$). The target is disease risk among untested patients ($R = 0$), conditional on their routine measurements. A representative documentation study records (H, C) for a sample of those patients without requiring scans. Collecting detailed assessments for this sample can be affordable even when obtaining them for every automated prediction is impractical.

We can fit $\widehat{p}_F(h, c) = \widehat{P}(Y = 1 \mid h, c, R = 1)$ on tested patients and the estimate $\widehat{\Pi}_0(dc \mid h)$ from the untested sample. SFLA then predicts $\int \widehat{p}_F(h, c)\, \widehat{\Pi}_0(dc \mid h)$ for a new patient with input h. Direct instead weights tested examples by the untested-to-tested density ratio on (H, C). Both require $Y \perp R \mid H, C$ and tested examples covering the target (H, C) values. An assessment that misses information used to choose the scan can violate the first assumption. A subgroup that is never tested violates the second. This example concerns disease status at assessment, before testing could change subsequent care.

Boeken et al. (2023) use selective CT testing to motivate privileged information. Their repeated regression targets the conditional mean via an outer regression; for binary outcomes this already defines a Bernoulli law, whereas continuous and joint prediction requires additional distributional structure. The same information pattern can use a trained probabilistic model or a frozen predictor queried with specified (h, c) values. Missingness handling alone supplies neither the untested sample nor its distribution of C.

Other ways to obtain auxiliary information. When replacing a selection model, transition records can contain its old features C and the new predictor's inputs H (Boeken et al., 2023). A representative sample that continues to measure both allows correction while subsequent individual predictions use H. Survey integration offers another established design: Kim et al. (2021) combine a selected outcome sample with a representative covariate survey. Their mass-imputation estimator targets population summaries. Here the additional requirement is to estimate a predictive distribution conditional on specified inference inputs.

A representative auxiliary sample must measure H and C together. Group totals or an unrelated archive of C cannot generally estimate $\Pi_0(C \mid H)$. Known sampling weights can recover the target composition from a designed subsample. If C is already available for each individual prediction, the corresponding (H, C) predictor is a richer-input comparator. Appendices B.2 and C.4 explain why retaining C only in labelled contexts does not resolve this population-information requirement.

A credit-score audit with selectively observed repayment. Suppose an auditor wants to compare a fixed credit score's prediction error across applicant groups G. Let H include reported income and credit history, and C the underwriter's manual assessment of supporting financial documents. For a specified loan contract, let Y indicate default if that loan is issued. Approved applicants provide (H, C, G, Y). A representative sample of all target applicants, including rejected applicants, provides (H, C, G) and the score $q(H)$, without requiring repayment outcomes. The audit can collect assessments and group annotations that the automated score does not use. Its target population includes both approval decisions.

The counterfactual repayment outcome must be well defined, with observed repayment agreeing with it for recipients of that contract. Identification then requires approval to be independent of this outcome conditional on (H, C, G), with positive approval probability on target support. These are substantive assumptions about lending, in addition to having the auxiliary records (Appendix G.3). In our notation, the resulting group-specific bridge and support are $Q_0(Y \mid H, C, G) = Q_1(Y \mid H, C, G)$ and $P_0^{H,C,G} \ll P_1^{H,C,G}$, where P_0 now describes all target applicants and P_1 those with observed repayment. Write $\Gamma_G = dP_0^{H,C,G}/dP_1^{H,C,G}$.

For $P_0(G = a) > 0$, a fixed predictor and an integrable loss,

$$\mathbb{E}_0[\ell\{q(H), Y\} \mid G = a] = \frac{\mathbb{E}_1[\Gamma_G\, \mathbf{1}\{G = a\}\, \ell\{q(H), Y\}]}{\mathbb{E}_1[\Gamma_G\, \mathbf{1}\{G = a\}]}. \tag{A.4}$$

This is the usual change-of-measure identity applied to the group-weighted loss and its normalizing probability. Conditioning the bridge on G is an additional assumption: stability after averaging

over groups does not imply stability within each group. Estimated weights require corresponding estimation and uncertainty analysis.

Operationally, estimate Γ_G from the labelled and auxiliary records and average the weighted observed losses within each group. An outcome-model alternative fits $\widehat{p}_F(h, c, g) = \widehat{P}(Y = 1 \mid h, c, g, R = 1)$. For the fixed score $q(h)$ and a loss such as Brier, its predicted loss is $\widehat{p}_F \ell(q, 1) + (1 - \widehat{p}_F)\ell(q, 0)$. Averaging this quantity over auxiliary applicants in each group estimates their prediction error, using sampling weights if required. This calculation integrates over the unknown binary outcome; squaring the difference between q and a fitted mean would omit outcome variation.

Coston et al. (2021) estimate predictive disparities under selective labels using an outcome model and study how those disparities vary across similarly accurate predictors. Our audit example makes the separation between adjustment records and prediction inputs explicit. It does not require a new fairness criterion or a generative model for binary outcomes. Chang & Wiens (2024) address a related clinical problem in which untested patients are assigned negative proxy labels and testing differs by group. Their EM classifier models that observation process. Applying our missing-outcome formulation there requires retaining which labels were verified, rather than treating proxy negatives as verified outcomes.

Group-wise prediction error can inform a fairness assessment, but the relevant outcome, decision rule and fairness criterion must also be specified (Fawkes et al., 2024). Improving average prediction error can still harm a subgroup.

B Proofs and score identities

The proofs answer three practical questions: what information identifies the target conditional distribution, how component error changes the gain from updating a mixture, and where recording weights may enter a fitted predictor. Conditional projection and change of measure are standard tools. Their role here is to distinguish errors that require different remedies: retaining information, correcting the population, or improving the predictive model. The controlled experiments in Appendices D and E test these distinctions.

B.1 Score geometry and conditional projection

We condition on fitting data independent of evaluation, prove the identities for the resulting fixed predictors, then average over fitting when needed. The moment conditions below ensure that all expectations exist.

Brier and energy excess risk. For vector Brier loss, let p be the true class-probability vector and q a forecast. Since $\mathbb{E}e_Y = p$, expansion gives

$$\mathbb{E}\|q - e_Y\|^2 - \mathbb{E}\|p - e_Y\|^2 = \|q\|^2 - 2\langle q, p\rangle + \|p\|^2 = \|q - p\|^2.$$

The embedding is the probability vector. Our binary vector-Brier convention is twice the squared error of the class-one probability. For energy, Euclidean distance is of negative type and admits a Hilbert embedding φ with $\|\varphi(x) - \varphi(y)\|^2 = \|x - y\|_2$ (Sejdinovic et al., 2013). Choose $\varphi(0) = 0$. Then $\mathbb{E}\|\varphi(X)\|^2 = \mathbb{E}\|X\|_2$, so finite first moments give a well-defined mean embedding $\mu(P) = \mathbb{E}_P\varphi(X)$. For mutually independent $X, X' \sim Q$ and $Y, Y' \sim P$,

$$\begin{aligned} D_{\mathrm{ES}}(P, Q) &= \mathbb{E}\|X - Y\|_2 - \tfrac{1}{2}\mathbb{E}\|X - X'\|_2 - \tfrac{1}{2}\mathbb{E}\|Y - Y'\|_2 \\ &= \|\mu(Q)\|^2 + \|\mu(P)\|^2 - 2\langle \mu(Q), \mu(P)\rangle \\ &= \|\mu(Q) - \mu(P)\|^2. \end{aligned}$$

The second moments of $\varphi(X)$ and $\varphi(Y)$ cancel, giving half the usual energy distance. Linearity of expectation gives $\mu(\sum_c \pi_c Q_c) = \sum_c \pi_c \mu(Q_c)$ for both scores.

Bayes risk and excess risk. For a proper score, conditioning on V shows that the true conditional law minimizes expected loss. Under the bridge, $Q_F = P_0(\cdot \mid H, C)$. Hence an independently fitted H-based forecast satisfies

$$\mathcal{R}_0(\widehat{Q}_H) = \mathcal{R}_0^\star(H, C) + \mathbb{E}_0 D_\ell(Q_F, \widehat{Q}_H).$$

Proposition 1 decomposes this excess risk into information loss and prediction error at fixed inputs. Taking $\widehat{Q}_H = Q_{0,H}$ gives $\mathcal{R}_0^\star(H) - \mathcal{R}_0^\star(H, C) = \Delta_{\text{info}}^\ell$. For binary vector Brier loss, the minimum at class-one probability η is $2\eta(1-\eta)$. In Table 2, the probabilities $0.1/0.9$ with (H, C) give risk 0.18, while $0.3/0.7$ with H give 0.42. Using the source probability 0.5 instead gives risk 0.50. The differences 0.24 and 0.08 are respectively Δ_{info} and Δ_{epi}.

Proof of Proposition 1. Write $m_F = \mu(Q_F)$ and $m_0 = \mathbb{E}_0[m_F \mid H]$. The bridge and the conditional-mixture identity imply $m_0 = \mu(Q_{0,H})$. Condition on the independently fitted prediction procedure. For its H-measurable prediction $\widehat{m}$, expand

$$\|m_F - \widehat{m}\|^2 = \|m_F - m_0\|^2 + \|m_0 - \widehat{m}\|^2 + 2\langle m_F - m_0, m_0 - \widehat{m}\rangle.$$

The inner product has expectation zero because

$$\mathbb{E}_0[\langle m_F - m_0, m_0 - \widehat{m}\rangle \mid H] = \langle \mathbb{E}_0[m_F - m_0 \mid H], m_0 - \widehat{m}\rangle = 0.$$

Square integrability makes the cross term integrable by Cauchy–Schwarz. Taking expectations proves the proposition. Substituting $\widehat{m} = \mu(Q_{1,H})$ gives the information–mismatch decomposition. For energy, Jensen's inequality gives $\|\mu(Q)\|^2 \le \int \|y\|_2 Q(dy)$ with the embedding chosen above. Integrable outcome and forecast first moments over deployment inputs therefore suffice. The forecast is the returned distribution, including any numerical sampling approximation.

Binary states and prediction inputs. For two state-conditional embeddings m_0^F, m_1^F, set $\delta = m_1^F - m_0^F$ and $d = \|\delta\|^2$. At a fixed query, $m_r = m_0^F + p_r\delta$. Hence

$$\begin{aligned}
\Delta_{\text{epi}} &= \|m_1 - m_0\|^2 = (p_1 - p_0)^2 d, \\
\Delta_{\text{info}} &= (1-p_0)\|m_0^F - m_0\|^2 + p_0\|m_1^F - m_0\|^2 \\
&= (1-p_0)p_0^2 d + p_0(1-p_0)^2 d = p_0(1-p_0)d.
\end{aligned}$$

Using the available inference inputs H, the source and deployment optima are $Q_{1,H}$ and $Q_{0,H}$, with excess risks $\Delta_{\text{info}} + \Delta_{\text{epi}}$ and Δ_{info}. Using (H, C), both are Q_F under the bridge and support. Relative to $\mathcal{R}_0^\star(H, C)$, the ideal predictors therefore have the following excess risks:

Prediction inputs	Source conditional	Deployment conditional
H	$\Delta_{\text{info}}^\ell + \Delta_{\text{epi}}^\ell$	$\Delta_{\text{info}}^\ell$
(H, C)	0	0

Integration retains any context dependence in $d(h)$. A recovery fraction is meaningful only when this oracle opportunity is positive and accurately estimated. The experiments report raw paired risks for zero-opportunity controls.

B.2 Indistinguishable environments

Two populations can provide identical information for fitting, selection and prediction while requiring different forecasts. The next result bounds average excess risk across them, giving the bound in Section 4.1 and its continuous-output analogue.

Proposition B.1 (Indistinguishable prediction populations)**.** *Suppose two populations have the same marginal law of H and give a learning procedure identically distributed observable inputs jointly with H. Evaluation outcomes are conditionally independent of these inputs and the procedure's randomness given H. Couple the observable inputs and randomness, so its forecast $\widehat{Q}(\cdot \mid H)$ is shared. Write $Q_\pm(\cdot \mid H)$ for the true conditionals and $m_\pm = \mu(Q_\pm)$, $\widehat{m} = \mu(\widehat{Q})$. For the Brier or energy embeddings, with finite expected squared distances,*

$$\frac{1}{2}\sum_{r\in\{+,-\}} \mathbb{E}_H D_\ell(Q_r, \widehat{Q}) = \mathbb{E}_H \left\| \widehat{m} - \frac{m_+ + m_-}{2} \right\|^2 + \frac{1}{4}\mathbb{E}_H D_\ell(Q_+, Q_-). \tag{B.1}$$

The final term lower-bounds average excess risk above each population's own Bayes risk. Expectations also average over fitting and prediction randomness.

Proof. Set $\bar{m} = (m_+ + m_-)/2$ and $a = (m_+ - m_-)/2$. The two errors are $(\widehat{m} - \bar{m}) - a$ and $(\widehat{m} - \bar{m}) + a$. Expanding their squared norms and averaging cancels the cross terms, leaving $\|\widehat{m} - \bar{m}\|^2 + \|a\|^2$. Integrating proves the identity. Nonnegativity of its first term gives the bound. The coupling preserves each procedure's marginal risk, so the argument includes randomized training and sampling. □

Pretraining and missing inputs. PFNs approximate posterior prediction under a distribution of training tasks (Müller et al., 2022). Let $\mathcal{O}$ collect the labelled context, visible inference inputs and missingness indicators. For an unrestricted probabilistic predictor trained with a strictly proper score, the population optimum is $q^\star_{\mathrm{pre}}(dy \mid \mathcal{O}) = P_{\mathrm{pre}}(dy \mid \mathcal{O})$. This follows by minimizing conditional expected score at each $\mathcal{O}$. A fitted network approximates this optimum through its architecture, training data and optimization. With a finite context, the posterior prediction also averages over uncertainty about the task.

Within a fixed training task, let $J = 1$ mean that C is artificially hidden. If $Y \perp J \mid (H, C)$, total probability gives, wherever the conditionals are supported,

$$P_{\mathrm{tr}}(dy \mid h, J = 1) = \int P_{\mathrm{tr}}(dy \mid h, c)\, P_{\mathrm{tr}}(dc \mid h, J = 1).$$

The conditional independence replaces $P_{\mathrm{tr}}(dy \mid h, c, J = 1)$ by $P_{\mathrm{tr}}(dy \mid h, c)$ in the integral. If also $J \perp C \mid H$, the mixing distribution is $P_{\mathrm{tr}}(dc \mid h)$. Thus masking can teach averaging over a missing state. The training population and masking rule determine which averaging is optimal. Missingness patterns can themselves be informative when these independences do not hold.

Deployment prediction requires the state distribution $\Pi_0(dc \mid h)$ under the bridge. Even an exact pretrained predictor faces Proposition B.1 when its complete inputs are identical in two deployments. Optimality averaged over a pretraining prior does not imply optimality in every compatible deployment. Representative auxiliary records can distinguish them and allow a joint learner to infer the mixture internally. SFLA estimates these probabilities explicitly and combines them with the state-conditional predictions. Our fixed-model comparisons evaluate inference interfaces and recording correction. Isolating pretraining or masking effects would require controlled changes to training.

Missing recording states in pretrained predictors. In the binary construction, $P_\pm(Y = 1 \mid X) = 1/2 \pm X/5$ and $X = \pm 1$. Both Bayes risks are $2(0.3)(0.7) = 0.42$. Writing $q(H)$ for the predicted class-one probability gives

$$\frac{\mathcal{R}_+(q) + \mathcal{R}_-(q)}{2} = 0.50 + 2\mathbb{E}_H(q(H) - 0.5)^2 \geq 0.50. \tag{B.2}$$

Thus at least one direction has excess risk at least 0.08. A predictor may match one direction, so this is a paired bound, not a separate 0.50 lower bound in each population. Source contexts can retain (H, C, Y) and inference inputs can explicitly mark C missing: neither distinguishes the two deployments. The bound applies to each interface in Appendix C.4, including one that retains C. Auxiliary (H, C) records distinguish the deployments and permit correction. This is additional population information; no individual query state is required. With exact components, pooled deployment weights still give $q = 0.5$ because both marginal state probabilities equal $1/2$.

Continuous outputs with unchanged coordinate marginals. For an analytical example, take $X = \pm 1$ equiprobably and $P_\pm(C = 1 \mid X) = 1/2 \pm X/4$. Let $Q^{(c)} = N(0, \Sigma_c)$ on two response coordinates, with unit variances and correlation $(-1)^c\rho$, where $0 < \rho < 1$. Every $Y_j \mid X$ is $N(0, 1)$ in both deployments. Their unconditional joint response laws also agree, since both marginal state probabilities equal $1/2$. Yet $\mathrm{Cov}_\pm(Y_1, Y_2 \mid X) = \mp\rho X/2$. Writing $d = D_{\mathrm{ES}}(Q^{(0)}, Q^{(1)}) > 0$, mixture linearity gives $D_{\mathrm{ES}}(Q_+, Q_-) = d/4$. Equation (B.1) therefore bounds average excess energy by $d/16$ from below. The equal-weight source mixture attains this bound; exact conditional SFLA attains the Bayes risk given H in each population. Coordinatewise CRPS has zero correction opportunity because its target marginals agree. The bound applies equally to generators trained from scratch and frozen predictors. Weighted flow fitting changes the information used in learning, as derived in Appendix B.8. Appendix D.3 tests the distinction between unchanged scalar marginals and changed joint dependence with fitted generators under a separate recording and query design.

B.3 Global class corrections

Consider an original binary classifier with $0 < q_B(1 \mid h) < 1$ and a positive class multiplier (r_0, r_1) that is constant over a deployment batch. Its normalized prediction satisfies

$$q_r(y \mid h) = \frac{r_y q_B(y \mid h)}{\sum_k r_k q_B(k \mid h)}, \qquad \operatorname{logit} q_r(1 \mid h) = \operatorname{logit} q_B(1 \mid h) + \log(r_1/r_0).$$

It represents an interior target probability $\eta_0(h)$ exactly if and only if $\operatorname{logit}\eta_0(h) - \operatorname{logit} q_B(1 \mid h)$ is constant almost everywhere. Necessity follows from the displayed identity. Choosing the corresponding ratio proves sufficiency. The multiclass version adds a constant to each pairwise log-odds ratio. In the binary case, this operation preserves the original classifier's ranking within the batch.

This characterizes the batch DistPFN variants we compare. The published method also considers per-instance adjustments (Lee, 2026). Even with auxiliary (H, C) data, a global multiplier remains restricted to a constant log-odds change. Its failure is therefore a model restriction, distinct from the information bound above.

Under classical label shift, stable $P(H \mid Y)$ gives a global class-prior ratio through Bayes' rule. Our bridge instead holds at (H, C) and gives the conditional mixture in Equation (2). For the unchanged-marginals experiment, the exact source predictor is $1/2$ while the target is $1/2 + 0.8\kappa X$. No constant odds correction can recover this target. SFLA can do so by retaining the association of X and C in its mixture probabilities.

For the exact source predictor, a global class adjustment produces a constant class-one probability q. Since $P_0(Y = 1) = 1/2$ in either deployment, its vector-Brier risk is

$$\mathcal{R}_0(q) = q^2 + (1 - q)^2 = 0.50 + 2(q - 0.5)^2.$$

The best adjustment therefore retains $q = 0.5$. This explains the equal 0.50 bars in Table 2: both retain the source mismatch $\Delta_{\text{epi}} = 0.08$ above the H-only optimum 0.42. Any other constant adjustment adds $2(q - 0.5)^2$ to that risk.

B.4 Signed gains of a fixed component library

Use the notation of Proposition 2. At target probability $p + x$, write $m_p = (1 - p)\mu_0 + p\mu_1$ and $\widehat{m}_p = (1 - p)\widehat{\mu}_0 + p\widehat{\mu}_1$. The true target embedding is $m_p + x\delta$, while the two forecasts are $\widehat{m}_p$ and $\widehat{m}_p + x\widehat{\delta}$. Since $\varepsilon = \widehat{m}_p - m_p$, their target errors are $\varepsilon - x\delta$ and $\varepsilon + x(\widehat{\delta} - \delta)$. Bayes risk cancels in the risk difference. Expanding both squared errors gives

$$\begin{aligned} G(x) &= \|\varepsilon\|^2 - 2x\langle\varepsilon, \delta\rangle + x^2\|\delta\|^2 \\ &\quad - \|\varepsilon\|^2 - 2x\langle\varepsilon, \widehat{\delta} - \delta\rangle - x^2\|\widehat{\delta} - \delta\|^2 \\ &= -2x\langle\varepsilon, \widehat{\delta}\rangle + x^2\{\|\delta\|^2 - \|\widehat{\delta} - \delta\|^2\}. \end{aligned}$$

Equivalently,

$$G(x) = -2x\langle\varepsilon, \widehat{\delta}\rangle + x^2\{2\langle\delta, \widehat{\delta}\rangle - \|\widehat{\delta}\|^2\},$$

because $\|\widehat{\delta} - \delta\|^2 = \|\widehat{\delta}\|^2 + \|\delta\|^2 - 2\langle\delta, \widehat{\delta}\rangle$. If $G(x) = ax + bx^2$ and both shifts are supported, the symmetric average $\{G(x) + G(-x)\}/2 = bx^2$ removes the source-error linear term. Exact components give $a = 0$ and $b = \|\delta\|^2$. Under an active outcome-null, $\delta = 0$ and $b = -\|\widehat{\delta}\|^2$: spurious component differences can make even exact reweighting harmful.

Integration uses one fixed context distribution. Changes in that law or post-reveal probabilities must remain inside the expectation. Estimated probabilities introduce additional error. Independent fits assess the response coefficients, while repeated shifts of one fit trace the same quadratic.

B.5 Errors in the assembled prediction

For categorical outcomes, let $g_c(h)$ and $\widehat{g}_c(h)$ be the true and fitted component probability vectors. Define $a = \sum_c \pi_0(c \mid h)(\widehat{g}_c - g_c)$ and $b = \sum_c(\widehat{\pi}_0(c \mid h) - \pi_0(c \mid h))\widehat{g}_c$. The true deployment prediction is $g_{\text{dep}} = \sum_c \pi_0 g_c$ and the fitted prediction is $\widehat{g}_{\text{dep}} = \sum_c \widehat{\pi}_0\widehat{g}_c$. Vector Brier risk then satisfies

$$\mathcal{R}_0(\widehat{g}_{\text{dep}}) - \mathcal{R}_0(g_{\text{dep}}) = \underbrace{\mathbb{E}_0\|a\|^2}_{\text{component error}} + \underbrace{\mathbb{E}_0\|b\|^2}_{\text{probability error}} + \underbrace{2\mathbb{E}_0\langle a, b\rangle}_{\text{signed interaction}}. \tag{B.3}$$

The following embedding proof also covers energy scores. With true and fitted component embeddings $m_c, \widehat{m}_c$, define $a = \sum_c \pi_c(\widehat{m}_c - m_c)$ and $b = \sum_c(\widehat{\pi}_c - \pi_c)\widehat{m}_c$. Add and subtract $\sum_c \pi_c\widehat{m}_c$ to obtain

$$\sum_c \widehat{\pi}_c\widehat{m}_c - \sum_c \pi_c m_c = a + b.$$

Thus its excess risk is $\mathbb{E}_0\|a\|^2 + \mathbb{E}_0\|b\|^2 + 2\mathbb{E}_0\langle a, b\rangle$, proving Equation (B.3). If the numerical sampler has additional embedding error n, replace $a + b$ by $a + b + n$. The expansion adds $\mathbb{E}_0\|n\|^2$ and $2\mathbb{E}_0\langle a, n\rangle + 2\mathbb{E}_0\langle b, n\rangle$. The cross terms can have either sign. Total excess risk therefore depends on the interactions among component, probability and numerical errors.

Blending complete predictions. For the whole-law blend $Q_\lambda = (1-\lambda)B + \lambda F$, $0 \le \lambda \le 1$, linearity of the embedding gives $\mu(Q_\lambda) = \mu(B) + \lambda d$, where $d = \mu(F) - \mu(B)$. Writing $e_B = \mu(B) - \mu(Q_0)$, expand

$$R(B) - R(Q_\lambda) = -2\lambda\mathbb{E}_0\langle e_B, d\rangle - \lambda^2\mathbb{E}_0\|d\|^2.$$

At $\lambda = 1$ this equals $g = R(B) - R(F)$, so $-2\mathbb{E}_0\langle e_B, d\rangle = g + \kappa$, with $\kappa = \mathbb{E}_0\|d\|^2$. Substitution proves

$$R(B) - R(Q_\lambda) = \lambda\{R(B) - R(F)\} + \lambda(1-\lambda)\mathbb{E}_0 D_\ell(B_H, F_H).$$

Writing $g = R(B) - R(F)$ and $\kappa = \mathbb{E}_0 D_\ell(B_H, F_H)$, some positive fraction improves risk exactly when $g + \kappa > 0$, except for degenerate zero-displacement cases. In particular, a harmful full update with $-\kappa < g < 0$ admits a beneficial smaller update. If $g \le -\kappa$, no positive fraction helps. The nonnegative diversity term can improve risk even without recording shift. Attributing a blend's gain to population correction therefore requires a separate comparison that holds this averaging effect fixed. Appendix B.10 gives the optimal fraction and the information needed to estimate it.

B.6 What the controls test

The controls remove different reasons for adaptation. When recording information is unrelated to the response given the available inputs, $Y_T \perp C \mid H$, the state-conditional distributions are identical. The recording-state probabilities can still differ between populations. With no recording shift, source and deployment distributions instead agree. Using the same probabilities in a fixed library then makes its two assemblies identical. A third control preserves each coordinate's conditional distribution while allowing their joint dependence to change. The Gaussian example in Appendix D.3.1 has zero marginal-CRPS opportunity and positive joint-energy opportunity by construction.

Weighted fitting can still matter when the conditional law is invariant. Where the ratios exist, $\Gamma(h, c) = \{dP_0^H/dP_1^H\}(h)\pi_0(c \mid h)/\pi_1(c \mid h)$. A change in the distribution of H can change the best restricted model despite an invariant Bayes conditional. For example, $Y = H \in \{0, 1\}$ has the same conditional law when $P(H = 1)$ changes from 0.1 to 0.9, but the best constant probability changes accordingly. Conversely, a fitted library can have different components under a true state null. Its mixture displacement measures the sensitivity of the fitted prediction. Population opportunity additionally depends on the true conditional laws. Consequently, a small fitted update or a high state-classification AUC is insufficient to determine that opportunity.

B.7 Bias and variance of mixture estimation

Fix fitted binary-state component probability vectors $\widehat{q}_0(h)$ and $\widehat{q}_1(h)$, independently of the auxiliary sample. Write $\widehat{d} = \widehat{q}_1 - \widehat{q}_0$ and $\varepsilon = \widehat{q}_0 + p\widehat{d} - q_0^\star$, where $p = P_0(C = 1 \mid H = h)$ and $q_0^\star$ is the true deployment class-probability vector. If a conditional probability estimator has bias b and variance v, then

$$\mathbb{E}_{\text{aux}}\|\varepsilon + (\widehat{p} - p)\widehat{d}\|^2 = \|\varepsilon\|^2 + 2b\langle\varepsilon, \widehat{d}\rangle + (b^2 + v)\|\widehat{d}\|^2. \tag{B.4}$$

Subtracting the true vector from the assembled forecast gives $\widehat{q}_0 + \widehat{p}\widehat{d} - q_0^\star = \varepsilon + (\widehat{p} - p)\widehat{d}$. Expand the squared norm and use $\mathbb{E}(\widehat{p} - p) = b$ and $\mathbb{E}(\widehat{p} - p)^2 = b^2 + v$. This proves the expected excess vector-Brier risk conditional on the fitted components. For a stratum containing a fixed number n of independent auxiliary observations, $K \sim \text{Bin}(n, p)$ and the half-count estimator $\widehat{p} = (K + 1/2)/(n + 1)$ give

$$b = \frac{1/2 - p}{n + 1}, \qquad v = \frac{np(1-p)}{(n+1)^2}.$$

These follow from the binomial mean and variance. For random stratum size, condition on its size before averaging. The independent auxiliary sample makes ε and $\widehat{d}$ fixed in this calculation. Increasing n removes probability-estimation error but leaves component error $\|\varepsilon\|^2$. Smoothing bias can interact with that error. Integration gives the population result. Continuous-context probability estimators require their corresponding bias and variance.

Expected gain with an estimated probability. Mixture estimation can outweigh a weak shift. Conditional on a fixed library and exact source probability, an independent, unbiased target-probability estimate with variance v gives $\mathbb{E}_{\text{aux}} G_{\text{estimated}}(x) = G(x) - v\|\widehat{\delta}\|^2$. The same expansion holds for any of the mixture-linear Hilbert embeddings above. Condition on a fixed prediction input and fitted library, and retain the exact source probability p of Proposition 2. Write $p_0 = p + x$, $\eta = \widehat{p}_0 - p_0$ and $r = \varepsilon + x(\widehat{\delta} - \delta)$. The updated forecast has error $r + \eta\widehat{\delta}$. If the independent auxiliary estimator has conditional bias b and variance v, subtracting its expected squared error from that of source assembly gives

$$\mathbb{E}_{\text{aux}} G_{\text{estimated}}(x) = G(x) - 2b\langle r, \widehat{\delta}\rangle - (b^2 + v)\|\widehat{\delta}\|^2. \tag{B.5}$$

Indeed, $\mathbb{E}\eta = b$ and $\mathbb{E}\eta^2 = b^2 + v$ in $\|r + \eta\widehat{\delta}\|^2 = \|r\|^2 + 2\eta\langle r, \widehat{\delta}\rangle + \eta^2\|\widehat{\delta}\|^2$. For an unbiased estimate, only the variance penalty remains. With exact components, $r = 0$, $d = \|\delta\|^2$ and a raw frequency $\widehat{p}_0 = K/n_0$, $K \sim \text{Bin}(n_0, p_0)$, this reduces to

$$\mathbb{E}_{\text{aux}} G_{\text{estimated}} = d\left\{(p_0 - p)^2 - \frac{p_0(1 - p_0)}{n_0}\right\}.$$

For $d > 0$ and nonzero shift, positive expected gain therefore requires $n_0 > p_0(1 - p_0)/(p_0 - p)^2$ in this idealized frequency model. This is a condition on average prediction risk over auxiliary samples. A high-probability guarantee or a selection rule would require further analysis. Estimated source probabilities, shared fitting data, smoothing and clipping require their corresponding error and covariance terms. With a varying input H, apply the conditional identity before integrating over deployment inputs. Continuous-context classifiers generally have both bias and variance.

B.8 Change of measure and conditional flow matching

The bridge factors both joint laws as $P_r^{H,C}(dh, dc)Q_F(dy_T \mid h, c)$. Assume $P_0^{H,C} \ll P_1^{H,C}$ and let $\Gamma = dP_0^{H,C}/dP_1^{H,C}$. For any integrable f,

$$\begin{aligned}\mathbb{E}_1[\Gamma f] &= \int \Gamma(h, c)\left\{\int f(h, c, y)Q_F(dy \mid h, c)\right\} P_1^{H,C}(dh, dc)\\ &= \int f(h, c, y)Q_F(dy \mid h, c)P_0^{H,C}(dh, dc) = \mathbb{E}_0 f.\end{aligned}$$

Appending the same independent source-noise and time laws to both populations preserves this identity. With $U = Y_T - Z_0$ and $Z_t = (1 - t)Z_0 + tY_T$, exact weighted flow matching therefore has objective

$$\mathbb{E}_1[\Gamma\|v(Z_t, t, H) - U\|^2] = \mathbb{E}_0\|v(Z_t, t, H) - U\|^2.$$

For this squared regression argument, assume $\mathbb{E}_0\|U\|^2 < \infty$, a stronger moment requirement than that of energy scoring alone.

For $Z_t = (1 - t)Z_0 + tY_T$, squared-loss regression has minimizer

$$v_r(z, t, h) = \mathbb{E}_r[Y_T - Z_0 \mid Z_t = z, t, H = h].$$

The regression identity follows by writing $U = v_r + (U - v_r)$ and using $\mathbb{E}_r[U - v_r \mid Z_t, t, H] = 0$ to remove the cross term. For any other field v, its objective exceeds that of v_r by $\mathbb{E}_r\|v(Z_t, t, H) - v_r(Z_t, t, H)\|^2$. Conditioning on (H, C), the common outcome distribution and interpolant construction give a common population minimizer $v_F(z, t, h, c)$. Conditioning on the available inference inputs H, the conditional average over C may change between populations. This shows that the population optimum with (H, C) is unchanged. Finite fitting and estimated weights can still change the fitted predictor.

To see the marginal-flow relation, take a smooth, compactly supported test function ψ and justify differentiation under the expectation. At a fixed conditioning context,

$$\frac{d}{dt}\mathbb{E}_r\psi(Z_t) = \mathbb{E}_r\langle\nabla\psi(Z_t), U\rangle = \mathbb{E}_r\langle\nabla\psi(Z_t), v_r(Z_t, t)\rangle.$$

This is the weak continuity equation for the interpolant marginals. Under the regularity and uniqueness conditions for the ODE and its continuity equation, integrating the field transports the source along these marginals (Lipman et al., 2023). For a singular terminal distribution, endpoint regularity needs

separate justification. An atomic endpoint is interpreted as a $t \uparrow 1$ limit. A learned field and numerical solver introduce separate approximations. Velocity regression fits the field through a regression surrogate. Likelihood and proper-score evaluation instead concern the induced endpoint distribution. DT-FM evaluates the distribution returned by its solver. Appendix F.3 also compares selection by held-out flow-matching loss.

Mixtures of fields. When interpolant densities exist, the coarse population field is

$$v_r(z,t,h) = \sum_c \pi_{r,t}(c \mid z,h) v_F(z,t,h,c), \qquad \pi_{r,t}(c \mid z,h) = \frac{\pi_r(c \mid h) p_t(z \mid h,c)}{\sum_d \pi_r(d \mid h) p_t(z \mid h,d)}.$$

This follows by conditioning the velocity target first on C and then on (Z_t, H). The responsibilities depend on the current position and time. Fixed initial weights $\pi_r(c \mid h)$ therefore do not generally produce the coarse field. SFLA avoids this calculation by drawing one state and sampling that component's entire endpoint law.

Conditioning an assembled law. Suppose a compatible joint component admits the factorization $q_c(y_O, y_T \mid h^-, a) = q_{c,O}(y_O \mid h^-, a) q_c(y_T \mid y_O, h^-, a)$. Sum this factorization over states, then divide by the mixture marginal density of y_O. On its positive support,

$$q_r(y_T \mid h^-, a, y_O) = \sum_c \frac{\pi_r^-(c \mid h^-, a) q_{c,O}(y_O \mid h^-, a)}{\sum_d \pi_r^-(d \mid h^-, a) q_{d,O}(y_O \mid h^-, a)} q_c(y_T \mid y_O, h^-, a).$$

This proves the posterior update in Equation (C.7). Both the weights and the component conditionals may change when values are revealed. A state-probability estimator that already conditions on those values provides the posterior in this formula and needs no further update.

B.9 Sensitivity to a violated bridge

Suppose the transported source conditional has embedding m_1 but the true deployment conditional is $m_1 + b$, with $\mathbb{E}_0\|b\|^2 \le \epsilon^2$. For two fixed forecasts with embeddings u, v, the error in their source-transported risk difference is $-2\mathbb{E}_0\langle b, u - v\rangle$: subtract $\|u - m_1\|^2 - \|v - m_1\|^2$ from $\|u - m_1 - b\|^2 - \|v - m_1 - b\|^2$ and cancel $\|b\|^2$. Cauchy–Schwarz bounds its absolute value by $2\epsilon\{\mathbb{E}_0\|u-v\|^2\}^{1/2}$. The bound is indexed by an assumed bridge error ϵ, which the imposed-recording experiments do not identify for a natural application. Appendix E.4 evaluates this identity for fixed learned forecasts under outcome-law departures that preserve everything observed by the learner.

B.10 Partial updates and context averaging

For the blend in Appendix B.5, differentiating $G(\alpha) = \alpha(g + \kappa) - \alpha^2\kappa$ gives the optimal fraction $\mathrm{clip}((g + \kappa)/(2\kappa), 0, 1)$ when $\kappa > 0$. If $\kappa = 0$, the predictions agree almost surely and all fractions tie. Predictions and deployment inputs identify κ without outcomes. Estimating g still requires permitted labelled validation and its population correction.

For predictions from m individual contexts, write $\overline{\mu} = m^{-1}\sum_b \mu_b$ and $V = \mathbb{E}_0 m^{-1}\sum_b \|\mu_b - \overline{\mu}\|^2$. The finite-mixture identity

$$m^{-1}\sum_b R_0(Q_b) - R_0(\overline{Q}) = V$$

follows by the same squared-distance expansion, without independence of contexts. Explicitly, for each evaluation case, write $\mu_b - m_0 = (\mu_b - \overline{\mu}) + (\overline{\mu} - m_0)$. Average the squared norms over b. The cross term is zero because $\sum_b(\mu_b - \overline{\mu}) = 0$. Integrating gives the identity. Uniform-to-targeted ensemble gain is consequently the change in average individual-context risk plus $V_D - V_U$. Equal numbers of contexts do not hold this averaging benefit fixed. Individual forecasts are needed to estimate it. Averaged probabilities alone cannot recover the terms. The identities require averaging the predictive laws themselves. Averaging logits, quantiles or scores defines a different operation.

Mean prediction and context variance. Condition on the fitted donor pool and the other fitted models. For constructor s, let $\mu_{s,b}(h)$ be the predictive embedding from context set b, $m_s(h) =$

$\mathbb{E}_K \mu_{s,b}(h)$ and $\sigma_s^2(h) = \mathbb{E}_K \|\mu_{s,b}(h) - m_s(h)\|^2$. With L conditionally independent, identically distributed context sets and $\mu_0(h) = \mu(Q_{0,H=h})$, the score geometry above gives

$$\mathbb{E}_K\{R_0(\overline{Q}_s) - R_0(Q_{0,H})\} = \|m_s - \mu_0\|_0^2 + \frac{\mathbb{E}_0 \sigma_s^2(H)}{L}, \tag{B.6}$$

where $\|f\|_0^2 = \mathbb{E}_0 \|f(H)\|^2$. To prove it, expand $\overline{\mu}_s - \mu_0 = (m_s - \mu_0) + L^{-1} \sum_b (\mu_{s,b} - m_s)$. The centered terms have zero conditional mean, so their inner products with the first term vanish. Conditional independence removes the cross-context terms, leaving L identical variances divided by L^2. For dependent context sets, cross-context covariances divided by L^2 remain. Independently randomized contexts can still share the fixed donor pool. For CRPS, centered CDFs provide the required $L^2(\mathbb{R})$ representation. Averaging more contexts reduces the randomization term. The error in their mean prediction remains and can persist even with correct donor frequencies.

A fixed context rule under a corrected loss.

Proposition B.2 (Risk correction with fixed contexts)**.** *Condition on data $\mathcal{D}$ used to construct the context pool, preprocessing and auxiliary models, independently of the queries evaluated below. Let $K(d\mathcal{B} \mid h, \mathcal{D})$ be a context rule using permitted inputs, with no query outcome or hidden query state. It may produce several context sets. Let $q_\psi(\cdot \mid h, \mathcal{B}, \mathcal{D})$ be the normalized forecast used at inference. Suppose the bridge holds, the deployment law of (H, C) is absolutely continuous with respect to its source law with derivative $\Gamma(H, C)$, and the same K and prediction pipeline are used in both populations. For every fixed ψ with integrable loss,*

$$\mathbb{E}_1[\Gamma(H,C)\mathbb{E}_K \ell\{q_\psi(\cdot \mid H, \mathcal{B}, \mathcal{D}), Y_T\}] = \mathbb{E}_0 \mathbb{E}_K \ell\{q_\psi(\cdot \mid H, \mathcal{B}, \mathcal{D}), Y_T\}. \tag{B.7}$$

Proof. After conditioning on $\mathcal{D}$, integrate over K and write the resulting loss as $g_\psi(h, y)$. It is measurable in the permitted query and response. The bridge makes $\mathbb{E}[g_\psi(H, Y_T) \mid H, C]$ the same in the two populations. Changing measure on (H, C) by Γ gives the displayed identity. Integrability justifies the iterated expectations. □

The identity applies to new queries conditional on the fitted models and context pool, and permits uniform contexts with weighted learning. The loss must evaluate the prediction used at inference: averaging scores, averaging quantiles and scoring a mixture give different objectives. The scoring rule must also distinguish an empirical forecast from draws used only to estimate the score of a continuous distribution.

Changing the context rule, preprocessing or output map changes the predictor being evaluated. We therefore evaluate the new pipeline on separate source queries with the same rule used at inference. The fixed-context identity does not equate risks across context constructors. With estimated recording weights, errors in the weights can affect both context choice and the loss estimate.

B.11 Assembly: targets and optimality

SFLA adapts a fitted distribution to a new population by changing its mixture probabilities. Reusing the components preserves their represented variation and dependence without fitting an outer outcome model. The operations below differ in what they average and which population they target:

Operation	What is averaged	Purpose
Source assembly	State-conditional distributions with source probabilities	Reconstruct source prediction from the component library
Deployment assembly (SFLA)	The same distributions with deployment probabilities	Predict deployment outcomes from the available inference inputs
Context averaging	Predictions from several labelled contexts	Define a component or a Context Direct predictor
Original/SFLA blend	Two complete predictive distributions	Choose how far to move from the original predictor

Exact-probability controls separate component fitting from probability estimation. Supplying the actual query state instead selects its state-conditional distribution and changes the prediction information.

Exact components. Under the bridge, the deployment law at fixed h is $Q_{0,H} = \int Q_F(\cdot \mid h, c)\,\Pi_0(dc \mid h)$. For a strictly proper score, this law uniquely minimizes conditional risk over all forecasts based on H, whenever the risks are finite. Indeed, conditional risk minus its Bayes value is the nonnegative score divergence $D_\ell(Q_{0,H}, Q)$, which vanishes only when $Q = Q_{0,H}$. Exact source assembly similarly minimizes source risk. The two mixtures coincide when their integrated laws agree. Equal conditional state distributions suffice, and different state distributions can also produce the same prediction when their component contributions cancel.

By change of measure, unrestricted proper-score Direct has the same deployment optimum (Appendix B.8). Flow Direct also requires a realizable field and correct integration. Finite fits and restricted predictors may miss this common optimum.

Repeated regression and distributional learning. Repeated regression (Boeken et al., 2023) fits a source conditional mean and regresses its predictions on deployment auxiliary records onto H. The following identity justifies its mean-prediction objective, including within a restricted regression class.

Proposition B.3 (Repeated-regression objective)**.** *Under* (1) *and finite second moments, let* $m(H, C) = \mathbb{E}_1[Y \mid H, C] = \mathbb{E}_0[Y \mid H, C]$. *For square-integrable* $g(H)$*,*

$$\mathbb{E}_0[(Y - g(H))^2] = \mathbb{E}_0[\mathrm{Var}_0(Y \mid H, C)] + \mathbb{E}_0[(m(H, C) - g(H))^2]. \tag{B.8}$$

The two fitting objectives have the same minimizers in any fixed class where those minimizers exist.

Proof. Expand $Y - g = (Y - m) + (m - g)$ conditional on (H, C). The cross term vanishes because $\mathbb{E}_0[Y - m \mid H, C] = 0$. Average over (H, C); the variance term is independent of g. □

With estimated models, error in both the conditional mean and its outer regression affects prediction. Auxiliary records must represent deployment or carry appropriate target weights. For binary outcomes, a probability in $[0, 1]$ defines a Bernoulli distribution. Calibration additionally requires $\mathbb{E}_0[Y \mid p(H)] = p(H)$; valid outputs or a lower risk than the original predictor do not establish this property.

Relation to distributional learning. The evaluated continuous repeated-regression baseline learns from pseudo-means. An extension could instead train a probabilistic predictor on the teacher's distributions or joint samples, preserving variation within each state. Under the bridge, an exact teacher and a strictly proper score, population fitting targets the deployment distribution when the predictor class contains it (Gneiting & Raftery, 2007). A fitted teacher targets the mixture of its estimated components. SFLA represents that mixture explicitly; a learned predictor would also depend on its output class and outer fitting error. Joint teacher samples must share the state across coordinates. Neither fitted procedure has an automatic calibration guarantee.

Coarsening recording information. A summary $\bar{C} = g(C)$ need not preserve the bridge. For almost every $(h, \bar{c})$ in each population, the tower property gives

$$P_r(dy_T \mid h, \bar{c}) = \int Q_F(dy_T \mid h, c)\, P_r(dc \mid h, \bar{c}).$$

The conditional measure is supported on $g(c) = \bar{c}$. Its source and deployment versions can assign different weights within that class, changing the resulting outcome law even when Q_F is stable. Coarsening preserves the bridge if the component law is constant within each class, or if the within-class conditional state distributions agree. A binary summary of richer collection information therefore requires its own bridge justification.

Predicting with a single state. For exact binary components at a fixed h, write $p = \pi_0(1 \mid h)$, $m_c = \mu(Q_F(\cdot \mid h, c))$ and $m_p = (1 - p)m_0 + pm_1$. Choosing the most probable state c^* gives conditional excess risk

$$D_\ell(Q_{0,H}, Q_F(\cdot \mid h, c^*)) = \min\{p, 1 - p\}^2 \|m_1 - m_0\|^2. \tag{B.9}$$

Indeed, $m_0 - m_p = -p(m_1 - m_0)$ and $m_1 - m_p = (1 - p)(m_1 - m_0)$; selecting the more probable state and squaring proves the formula. For $0 < p < 1$, this penalty is positive whenever

the components differ in the chosen embedding. In the binary experiment, $p = 1/4$ or $3/4$ and the squared component distance is 1.28, so the penalty is 0.08 and risk is 0.50, compared with 0.42 for exact assembly. The same formula holds for CRPS and joint energy, with their own component separations. Fitted components can change these risk rankings, as analyzed next.

Population probabilities and fitted predictions. With imperfect components, the prediction-optimal weights can differ from the population probabilities. Fix h, let m_c be the fitted component embeddings, m_* the true deployment embedding, $p = \pi_0(1 \mid h)$, $d = m_1 - m_0$ and $\varepsilon = m_0 + pd - m_*$. For $\|d\| > 0$, minimizing $\|m_0 + td - m_*\|^2$ over $0 \le t \le 1$ gives

$$t^* = \mathrm{clip}_{[0,1]}\left(p - \frac{\langle \varepsilon, d \rangle}{\|d\|^2}\right). \tag{B.10}$$

Differentiate the scalar quadratic and clip its unconstrained minimizer. An error along d can therefore be offset by changing the weight. If $d = 0$, all weights give the same embedding. Exact-probability controls measure the practical effect (Appendix D.1). Auxiliary (H, C) records identify p. Learning prediction-optimal weights additionally requires outcome information, such as separate source labels with valid recording weights.

Averaging predictions across contexts. With fixed mixture weights, averaging component distributions across context sets and then applying SFLA gives the same distribution as applying SFLA to each set and averaging the results. This follows from linearity of mixtures. Concatenating contexts in a nonlinear predictor or averaging logits or quantiles changes the operation. Appendix B.10 quantifies the score gain from averaging and from blending the original prediction with SFLA.

Scalar and joint outputs. Combining class probabilities or scalar CDFs implements the same mixture operation on different output spaces. For a joint block, one state draw followed by a full component draw implements the vector mixture. Independently mixing each coordinate instead gives the product of the mixture marginals. For the following covariance calculation, assume finite second moments and write $\mu_{cj} = \mathbb{E}[Y_j \mid H, C = c]$. The joint mixture has

$$\mathrm{Cov}(Y_j, Y_k \mid H) = \sum_c \pi_c \,\mathrm{Cov}(Y_j, Y_k \mid H, C = c) + \mathrm{Cov}_{C \sim \pi}(\mu_{Cj}, \mu_{Ck}).$$

This follows by conditioning the product moment on C and subtracting the product of the mixture means. Independent marginal sampling removes both terms. Sharing the state but sampling independently within it retains only the second. The completed-table experiment isolates this distinction. For a partial query, both weights and components must condition on the revealed values, as derived in Appendix B.8.

Consistency across prediction queries. Each evaluated query returns a normalized distribution. Interpreting these predictions as conditionals of one common joint distribution requires additional compatibility. A joint component model supplies this property when both components and mixture probabilities are conditioned on revealed values, as in Appendix B.8. Separately fitted conditional models need not satisfy it. A fixed sequential sampling order defines a joint distribution, but need not agree with predictions in another order or under other masks.

C Estimation and evaluation

Recording correction can use the interfaces already available in a predictor: model fitting, output heads, labelled contexts or component predictions. Table C.1 maps these interfaces to Direct and SFLA. We specify the estimators and evaluation rules below.

C.1 Data allocation and probability estimation

Source labels fit the predictors, and outcome-free auxiliary records estimate recording and state probabilities. A separate set of source labels selects the procedure or predicts its deployment gain before any deployment outcomes are used for evaluation. Each experiment's splits and uncertainty estimates follow the independent units summarized in Table C.2.

Table C.1: How the prediction interfaces implement recording correction. All routes use labelled source records and auxiliary population information. The individual prediction uses only H. Context targeting and weighted learning are evaluated separately.

Procedure	Required model access	How recording information enters
Full-model Direct	Trainable parameters	Weighted query loss
Head Direct	Predictions on separate labelled examples	Weighted loss for an output correction
Context Direct	Control of labelled demonstrations	Recording-weighted sampling of examples
SFLA	State-conditional probabilities, CDFs or joint samples	Deployment probabilities combine the components
Source validation	Candidate predictions on held-out labels	Weighted scores choose the returned predictor

For contextual recording probabilities, the specified estimator uses histogram gradient boosting on H and a one-hot encoding of C: 100 iterations, at most seven leaves, minimum leaf size 64, and L_2 penalty one. Auxiliary groups are divided into fitting, tuning and evaluation fractions $1/2, 1/4, 1/4$. The tuning split chooses a convex shrinkage toward the pooled fitting probability by Brier loss. Probabilities are floored at 0.001 before shrinkage. We retain the fitted models after tuning. When used for state probabilities $\pi_r(c \mid h)$, the predictors are H and the response is C.

Finite-state experiments use half-count or independent-frequency estimates. Deployment state probabilities account for the population change. The full ratios in Appendix A.4 are estimated before outcome fitting and selection, then held fixed.

Estimating general state distributions. For continuous or mixed C, fit a conditional density or sampler for $\Pi_e(\cdot \mid h)$ from auxiliary records. A vector of states can be factorized into successive categorical or continuous conditionals. Independent draws $c_b \sim \widehat{\Pi}_e(\cdot \mid h)$ yield the normalized approximation $B^{-1}\sum_b \widehat{Q}_F(\cdot \mid h, c_b)$, whose probability for any fixed event converges to the fitted integral by the law of large numbers. Drawing one state and then an outcome samples the mixture directly. Joint targets share that state. The reported experiments use finite states; richer states additionally require accurate state models and sufficient source coverage.

Support and weight concentration. For nonnegative weights with positive total, report $n_{\text{eff}} = (\sum_i w_i)^2 / \sum_i w_i^2$ separately for training and validation. Applying the same expression to group totals measures concentration across groups; uncertainty still accounts for within-group dependence. Effective sample size and probability floors describe the estimator's finite-sample behavior. They do not establish the bridge or population support.

C.2 Direct procedures

The full-model objective is

$$\widehat{L}_e(\theta) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\Gamma}_e(h_i, c_i)\mathbb{E}_{\mathcal{B}\sim K_e(\cdot\mid h_i,\mathcal{S})}\ell\{Q_{\theta,e}(\cdot \mid h_i, \mathcal{B}), y_{T,i}\}. \tag{C.1}$$

The descriptor $d_e(h)$ identifies the deployment for a shared model. A separately fitted model may omit it. For a frozen classifier, let $\widetilde{q}_e$ be clipped, renormalized backbone probabilities and $\widetilde{h}$ source-scaled features. Head Direct returns

$$g_{\psi,e}(h) = \operatorname{softmax}\left[\log \widetilde{q}_e(h) + r_\psi\{\widetilde{h}, \widetilde{q}_e(h), d_e(h)\}\right], \tag{C.2}$$

Context Direct instead returns

$$\widehat{Q}_e^{\text{ctx}}(\cdot \mid h) = \frac{1}{L}\sum_{l=1}^{L} f_\phi(\cdot \mid h; \mathcal{B}_{e,l}), \qquad (\mathcal{B}_{e,1}, \ldots, \mathcal{B}_{e,L}) \sim K_e(\cdot \mid h, \mathcal{S}). \tag{C.3}$$

Here $\mathcal{S}$ is the labelled source pool, f_ϕ the frozen predictor and K_e the context constructor specified below. Full-model Direct uses (C.1), with conditional flow-matching loss for flows. The classification head adds a residual to frozen log probabilities and recovers the clipped, normalized original when that residual is zero.

Scalar output heads. For a frozen scalar CDF F_h, divide probability rank into four equal intervals I_j and fit their mixture weights:

$$Q_{h,j} = \mathcal{L}\{F_h^{-1}(U) \mid U \in I_j\}, \quad U \sim \text{Uniform}(0,1), \qquad Q_{\alpha,h} = \sum_{j=1}^{4} \alpha_j Q_{h,j}, \quad \alpha \in \Delta^3. \tag{C.4}$$

We split atoms crossing a boundary by mass. Uniform weights recover the original prediction. One weight vector per response, backbone and condition minimizes the convex quadratic

$$\widehat{R}_w(\alpha) = \frac{1}{n} \sum_i w_i \ \text{CRPS}(Q_{\alpha,H_i}, Y_i), \qquad w_i = 1 \text{ or } \widehat{\Gamma}_i. \tag{C.5}$$

The loss scores the complete mixture, including cross-component terms. Under the bridge and correct ratios, weighting targets deployment risk while contexts and pretrained parameters remain fixed.

Sampling contexts. Each of L contexts contains m labelled source examples sampled independently with replacement, with probabilities proportional to the deployment ratio. Prediction averages the resulting L probability vectors or CDFs. The uniform control uses probabilities $1/n$ and matches the number of examples, contexts, query clusters and random uniforms. Repeated examples occupy context positions but supply no additional labels. Query labels are excluded from their contexts. Labels used to fit a head are disjoint from the context pool and count toward the total budget.

The general constructor K_e in Equation (C.3) includes dependent donor counts. The balanced rule starts from $a_i = mp_{e,i}$, retains $\lfloor a_i \rfloor$ copies, and rounds the fractional parts f_i. For two fractional entries i, j, set $u = \min(1 - f_i, f_j)$ and $d = \min(f_i, 1 - f_j)$. Replace (f_i, f_j) by $(f_i + u, f_j - u)$ with probability $d/(u + d)$, and by $(f_i - d, f_j + d)$ otherwise. Each step preserves their total and individual expectations in exact arithmetic and makes at least one entry integral. Repeating therefore gives $\sum_i N_i = m$, $\mathbb{E}N_i = mp_{e,i}$ and $N_i \in \{\lfloor a_i \rfloor, \lceil a_i \rceil\}$.

We randomize the order of fractional entries and final context positions, sharing random streams between paired procedures. With uniform weights and $m = n$, every donor appears once. Its matched expected frequencies do not imply the same context distribution as independent sampling, so the fixed-context risk identity does not order their prediction risks.

Context choice and query loss. Weighted query learning permits uniform or targeted contexts when learning and deployment share the context distribution, preprocessing, query companions and output rule (Proposition B.2). Validation scores the distribution actually used for prediction, including any averaging. A change to contexts or outputs therefore requires a new evaluation on separate source queries. Appendix F.2.1 compares source and weighted learning from fixed pretrained outputs.

Shared Direct. A shared model minimizes $\sum_e \nu_e \widehat{L}_e$ with descriptor $d_e(h) = \widehat{\pi}_e(\cdot \mid h)$ and fixed $\nu_e \geq 0$, $\sum_e \nu_e = 1$. Equal weighting of two compositions gives $\nu_e = 1/2$. The shared flow normalizes weights by their empirical mean within each composition. The head uses the unnormalized weighted mean.

Query and deployment weights. The query rule specifies which response coordinates to predict. Our fixed-block studies use one query; multi-query training samples supported queries with the specified frequencies. Recording weights correct the source-to-deployment difference for that query. If the query frequencies also change, their contribution belongs in the full density ratio on the query and observed inputs. A ratio already defined on this law receives no additional query-frequency factor. The bridge and support assumptions must hold for the resulting prediction problem.

Context matching. LoCalPFN (Thomas et al., 2024) combines retrieval and fine-tuning. We treat native and recording-corrected selection as separate procedures. Our CRUMB (Heredge et al., 2026) reimplementation uses random Fourier features and greedy subset selection to match visible query features. The recording-aware extension matches the estimated joint (H, C) law. Source- and deployment-targeted versions share feature maps, bandwidths, clusters and donor budgets, while receiving more auxiliary information than visible-feature CRUMB.

C.3 Assembly and conditional generation

Assembly estimates state-conditional components and mixes them with estimated population probabilities (Appendix B.11). Source assembly retains source probabilities to measure the effect of replacing the original predictor with a mixture of state-conditional predictions. A learned combination $Q_\beta = \sum_j \beta_j Q_j$ instead fits simplex coefficients by a proper score. These coefficients optimize prediction risk and need not estimate state probabilities.

SFLA fixes the component prediction rule, including its source pool, preprocessing, retrieval and output definition. Classifiers return

$$\widehat{g}_0(h) = \sum_c \widehat{\pi}_0(c \mid h)\widehat{g}(h, c). \tag{C.6}$$

Components may use separate state models or one shared state-conditioned model, including separate or state-labelled in-context examples. These choices affect estimation error and cost. A joint sample draws one state for the entire vector. Drawing a separate state for each coordinate changes the dependence.

The conditional flow draws independent Gaussian noise on the target coordinates and integrates its field with the observed coordinates held as conditioning inputs. A conditional Gaussian mixture instead updates its component responsibilities and Gaussian conditional means/covariances analytically. Both reuse a fitted model over supported queries. Holding coordinates fixed in an unconditional field generally gives a different process: the conditional field must be trained with the observed values as conditioning inputs.

When revealed values are informative about C, a compatible joint model gives

$$\pi_r(c \mid h^-, a, y_O) \propto \pi_r^-(c \mid h^-, a) q_{F,O}(y_O \mid h^-, c, a). \tag{C.7}$$

Normalize over supported states, then sample $q_F(y_T \mid h^-, c, a, y_O)$. Our query-conditioned flows use a separate statistical posterior model. Consistency across masks is therefore an additional requirement (Appendix B.11). Probabilities already conditional on Y_O receive no second likelihood update.

C.4 Distributional outputs from pretrained models

Missing recording-state inputs. Suppose labelled contexts contain an observed, varying C, but every inference input lacks it. Both the classification and regression implementations of TabICLv2 used here remove that column from context and inference arrays after preprocessing. TabDPT retains the context column and fills missing inference values with its source-column mean. Its optional indicators cover only columns missing during source fitting, so fully observed source C receives no indicator. In TabPFN package v9.0.0, the inspected v3 and v3.5 (Jäger et al., 2026) implementations retain the varying context column and encode missing entries with explicit indicators alongside imputed values.[2] Figure D.1 illustrates these interfaces and their classification comparison. We hold pretrained weights fixed and evaluate native prediction and recording correction separately (Appendix D.1.1).

The Original rows in Table 3 instead omit C from the labelled contexts as well as the inference inputs. They establish a common H-only reference across backbones. TabICLv2's native feature exclusion supplies the same variables to its predictor, although preprocessing and ensemble construction can differ from fitting on H alone. For TabDPT, retaining source C and mean-filling its inference values is a separate native procedure, whose performance is not measured by these Original rows. The recording process still determines which rows supply labels, and H may contain proxies for C.

For a nonlinear predictor, filling C with a mean generally differs from averaging its predictions: $f(h, \mathbb{E}[C]) \neq \mathbb{E}[f(h, C)]$. On the binary construction in Table 2, however, the oracle probability $0.1 + 0.8c$ is affine: evaluating it at the correct conditional deployment mean recovers the exact mixture. Native source-mean filling does not supply that deployment mean, and source observations at $C = 0, 1$ alone do not determine a fitted predictor's extension to fractional states. SFLA separates two estimation tasks: predicting within each supported state and estimating the deployment probabilities

[2] Inspected source versions: TabICLv2 `59a957c` (classifier, regressor), with column removal in its preprocessing; TabDPT `97e5494`; TabPFN v9.0.0 v3 and v3.5.

$\widehat{\pi}_0(c \mid h)$. It evaluates states explicitly before averaging, so the predictor need not infer these probabilities from a missing-value token. Training with missing inputs can teach averaging under the training population (Appendix B.2). Deployment probabilities still require identifying information. Proposition B.1 applies whenever the complete native inputs agree across deployments, including retained source states and explicit missingness indicators.

Applicability across model families. SFLA requires conditional outputs that can be combined: probability vectors for classes, CDFs for scalar outcomes, or a joint sampler for target vectors. For a mean-only predictor, distributional adaptation requires an additional output model. The Direct variants depend on the available interface: Context Direct controls demonstrations, Head Direct uses outputs on separate labelled examples, and Full-model Direct updates trainable parameters. For SFLA, explicitly supplying each supported state can also avoid exclusion of an entirely missing inference column, without observing the individual's state.

Additional interfaces illustrate the same distinction between representation and identification. Google TabFM retains source C and mean-fills missing entries. EXAONE Tabular (Eo et al., 2026) retains a separate missingness channel. Both inspected LimiX (LimiX Team, 2025) supervised filters remove a column missing throughout the inference batch.[3]

A joint learner given representative deployment (H, C) records could learn the mixture internally. Such a comparator would receive the same auxiliary information as SFLA, with source and deployment records distinguished and auxiliary outcomes unavailable. The native-input bound in Proposition B.1 concerns identical inputs across deployments. This richer comparison instead tests how effectively procedures use information that can identify the target distribution.

Scalar distributions. The TabICLv2 scalar adapter averages the 999 raw quantiles returned by its eight native ensemble members and sorts the resulting knots q_j at levels $a_j = j/1000$. Linear interpolation defines the interior quantile function. Its tails are

$$Q(a) = q_1 + \beta_L \log(a/a_1) \quad (a < a_1), \qquad Q(a) = q_{999} - \beta_R \log\{(1-a)/(1-a_{999})\} \quad (a > a_{999}).$$

Slopes are fitted by the release's log-space regression on twenty boundary quantiles, with bounds $[0.01, 100]$ in source-standardized target units. The adapter uses float64 and untruncated exponential tails. Repeated knots define atoms. These choices produce a normalized law with finite moments. Its mean includes the tail contribution and can differ from the native average-of-knots point prediction. Across states, assembly mixes CDFs:

$$\widehat{F}_0(z \mid h) = \sum_c \widehat{\pi}_0(c \mid h) \widehat{F}_c(z \mid h). \tag{C.8}$$

A state draw followed by an independent inverse-CDF draw samples this mixture. Averaging the components' quantiles would define another law.

TabDPT's full scalar output is a finite-support distribution. We combine its raw-scale CDFs, including when component supports differ, and include out-of-support outcomes in scoring. Native point and full distribution outputs have different ensemble aggregation and are identified separately. The iterative TabICLv2 imputer uses two conditional-update sweeps with independent restarts. We evaluate the joint distribution after those two sweeps. Establishing convergence to a stationary law would require a separate analysis. Multivariate imputation from scalar CDFs requires a joint construction, such as the sequential one below.

From scalar laws to a joint prediction. A fixed ordering $j_1, \ldots, j_m$ of a wholly missing block gives one explicit extension of a scalar predictor:

$$\widehat{Q}_c(dy \mid h) = \prod_{r=1}^{m} \widehat{Q}_{c,j_r}(dy_{j_r} \mid h, y_{j_1}, \ldots, y_{j_{r-1}}).$$

Source contexts use preceding coordinates as features. At inference, these features take the values generated at earlier steps. Iterated integration proves normalization when every scalar kernel is

[3] Pinned author implementations: TabFM `fbb6655`; EXAONE `8638e07`; LimiX `516bf39`, v1 and v2 filters.

normalized. Draw C once, then run this sequence to sample SFLA. Independent state draws or scalar laws conditioning only on (h, c) give different dependence structures.

The joint-TFM experiment uses this construction in a fixed order (Appendix D.3). Agreement across separately fitted orders or query masks requires the compatibility condition in Appendix B.11.

C.5 Scoring predictive samples

For iid draws $X_1, \dots, X_m \sim Q$, $m \geq 2$, the fair energy estimator is

$$\widehat{\mathrm{ES}}(Q, y) = \frac{1}{m} \sum_i \|X_i - y\| - \frac{1}{2m(m-1)} \sum_{i \neq j} \|X_i - X_j\|. \tag{C.9}$$

The first term estimates $\mathbb{E}_Q\|X - y\|$. Pairs with unequal indices are independent, so the second term estimates half the predictive diversity. Integrability gives unbiasedness. A denominator m^2 includes zero diagonal terms and underestimates diversity. Absolute distance gives CRPS for a scalar response. Euclidean energy scores the joint law. "Fair" corrects finite-ensemble bias (Ferro, 2014). Tables state source-fitted scales and any division by $\sqrt{|T|}$. Binary vector Brier ranges from zero to two. When the prediction itself is an empirical distribution, its exact score includes the diagonal terms.

For fixed weights and component sample sets whose vectors $(X_{1i}, \dots, X_{Ki})$ are iid across i, the mixture diversity estimator is

$$\sum_{c,d} \pi_c \pi_d \frac{1}{m(m-1)} \sum_{i \neq j} \|X_{ci} - X_{dj}\|.$$

Unequal indices again give independent draws, so its expectation equals the mixture's diversity. Excluding equal indices is necessary even when $c \neq d$: components may share simulation noise. Fitting a transformation to the same samples, such as subtracting their mean, can also couple unequal indices. Unbiasedness then needs a separate proof.

Coverage uses returned marginal quantiles and RMSE the mean of that law. Point methods are evaluated by point scores or explicitly as Dirac-law controls, whose diversity is zero. Joint-event Brier and completed-table products evaluate other functionals. Paired intervals use the stated independent unit: generated procedures, refits, stays or time groups. Coordinates, queries and predictive draws share the uncertainty of their fitted model.

C.6 Model selection and computational cost

Equation (6) scores the complete prediction on separate source labels, including context averaging and numerical sampling. The joint-association experiment minimizes validation energy. The common flow applications and complete-source benchmark choose the earliest checkpoint within one paired standard error of minimum risk. The frozen-model libraries specify their tie order in Appendix F.2.2. Held-out flow loss supplies a separate validation criterion, following related diffusion-loss stopping (Favero et al., 2025).

Checkpoint comparisons hold the sampler fixed, and sampler comparisons hold the checkpoint fixed. All selection uses source labels, with auxiliary deployment information where weights are required. Increasing the library can improve its best candidate while making selection less accurate (Appendix F.2.2). Pointwise intervals for individual contrasts do not give simultaneous guarantees for the selected predictor.

Cost includes training every candidate, estimating the correction, validation and inference. An early selected checkpoint still incurs the cost of the full training trajectory. Prediction on new queries includes context processing and sample generation, in addition to any mixture calculation. Pretraining is reported separately.

Algorithm 1 summarizes fitting, inference and source-based selection for the method in Section 3. The experiment-specific candidate sets and selection rules complete this specification.

Algorithm 1 Recording-aware prediction

Require: Labelled source (H, C, Y_T); unlabelled deployment sample (H, C); query h; separate source labels for output learning or selection

Fix a procedure, label allocation, candidate set and selection rule before evaluation (Table C.1). Require the bridge and source support; apply the procedure's eligibility checks without treating them as identification tests.
Estimate and freeze the required $\widehat{\Gamma}_e$ and $\widehat{\pi}_e$.
for each candidate procedure **do**
  if full-model or head Direct **then**
    Fit with the inference-time context rule and weighted source loss ((4) for flows). Shared Direct averages over specified compositions with descriptor $d_e(h)$. Retain matched unweighted controls.
  else if Context Direct **then**
    Draw contexts using K_e with source-row probabilities proportional to $\widehat{\Gamma}_e$; average the predictive distributions.
  else if SFLA **then**
    Fit and freeze state-conditional components. Retain source assembly with $\widehat{\pi}_1$; SFLA uses $\widehat{\pi}_e(c \mid h)$ with those same components. For joint samples, draw one state per block.
  else
    Construct the specified original, uniform-context or native-learning control.
  end if
end for
if the procedure learns output combinations **then**
  Freeze predictions and fit simplex mixture coefficients on separate labels by minimizing the specified score.
end if
if the procedure selects among candidates **then**
  Score each predictive distribution on separate source labels with deployment weights and a proper endpoint score; apply the prespecified rule.
end if
return Fixed or selected predictive distribution at h, without the query's actual state C.

C.7 Data sources and allocation

Table C.2 distinguishes measured-outcome tasks from studies that use public features with generated responses. Counts refer to the data splits. Each linked design gives eligibility and recording rules.

Sources and access. The linked UCI records for Protein (CASP), SGEMM, Tetouan, Gas Turbine, Appliances, Beijing, Household, Adult, Metro (Hogue, 2019) state CC BY 4.0 terms. California Housing is obtained through scikit-learn's StatLib distribution. The clinical study uses PhysioNet Challenge 2012 v1.0.0, whose files have open access under Open Data Commons Attribution v1.0.

C.8 Settings and recorded computation

Table C.3 records the settings for each comparison. The flows share a training configuration across tables. The completion and clinical studies have separate budgets and samplers.

Table C.2: Principal data allocations. Auxiliary rows contain only features and recording states; target outcomes are excluded. Exact group definitions and model-specific access appear with each experiment.

Data and use	Fitting and evaluation samples	Recording and uncertainty
California, Protein, SGEMM, Tetouan: measured completion	4,096 source rows (4,176 for Tetouan); 512 evaluation rows; three fits per table.	No recording shift. Two query sizes share each fit (F.4).
Gas Turbine and Appliances: measured scalar outcomes	2,048 fitting rows, 512 scaling rows and 256 source-validation rows per task.	Response-dependent state assignment and imposed recording; time groups condition on the fitted allocation (F.1.3).
Beijing: PM2.5 and PM10	2,048 fitting, 512 scaling, 4,096 auxiliary and 2,048 evaluation rows.	Station-based recording; 60 calendar-week groups. The later validation comparison adds 256 source labels (F.1.2).
Household: active and reactive power	2,048 fitting, 512 scaling, 4,096 auxiliary and 2,048 evaluation rows.	Submeter-based recording; no-shift control; 80 evaluation weeks from one household and one fit (F.1.2).
Adult: income classification	Matched comparison: 1,024 fitting and 512 validation labels; larger-context predictions are identified separately.	Two overlapping imposed assignments, each with a control with an unrelated recording state (F.6).
Metro: traffic volume	2,048 fitting, 512 scaling, 4,096 auxiliary and 2,048 evaluation rows.	Day-type recording; 99 represented evaluation weeks from one sensor and one fit per condition (F.1.2).
Electricity: hourly consumption	2,048 donor and 256 head labels; five or 64 validation clients; 114 evaluation clients.	Weekend-based recording with reversed and no-shift controls; client/calendar-block intervals over 43 weeks (F.2.2).
California: selectively recorded responses	4,352 source rows split among fitting, correction and two validation splits; 512 evaluation rows.	Longitude-based recording of income and housing age; one fit per model/condition (F.5).
PhysioNet Challenge 2012, set A	3,453 eligible stays; shared 689-stay evaluation set; four assignment-and-refit repetitions.	Imposed rank-based assignment for hematocrit/platelets; stay and refit uncertainty reported separately (F.7).

Table C.3: Settings for the principal comparisons. Clinical flows use a separate training and sampling budget from the completion benchmark.

Comparison	Settings
Complete-source DT-FM	32,768 updates; batch 128; endpoint selection on separate source cases.
Measured-application flows	Four width-256 residual blocks; 32,768 updates; batch 128; AdamW, learning rate 10^{-3}, weight decay 10^{-5}, gradient clip 2; 512 midpoint steps. Checkpoints at 256, 1,024, 8,192 and 32,768 updates; earliest eligible within one paired standard error of the best weighted validation energy (F.3).
Complete-source CFMI	75,000 updates; batch 64; native 100-step Euler and selected 400-step output.
Complete-source DiffPuter	Training-loss-best native checkpoint versus endpoint-selected retained epoch; native 50-step conditional sampler unchanged.
Recording estimator	Histogram boosting: 100 iterations, at most seven leaves, minimum leaf size 64, $L_2 = 1$; separate fitting, tuning and evaluation fractions $1/2, 1/4, 1/4$.
Context construction	Uniform and ratio-targeted sampling share context lengths and random uniforms; joint-information controls with 205 examples per context and larger application contexts are reported separately.
Scalar output	TabICLv2: 999 averaged/sorted quantiles, specified interpolation and tails. TabDPT: full finite-support CDF. Assembly mixes raw-scale CDFs.
TabPFN 3.5 applications	Eight ensembles; 2,048 examples per context and 512 scaling labels; unchanged float32 pretrained weights, float64 histogram conversion and aggregation; 128-draw fair CRPS.
Native TabICLv2 tuning	Pinball loss; learning rate 10^{-5}; at most 30 epochs; source-MSE patience eight; epoch zero eligible. The 30/100-epoch objective–stopping comparisons are separate controlled training configurations.
Clinical flow	Four width-256 residual blocks; 3,000 updates; raw/EMA checkpoints; 128 selection draws and 1,024 evaluation draws.
Scoring	Unequal-index energy and CRPS; source-fitted scales. Energy tables state normalization by target dimension; vector Brier ranges from zero to two.

D INFORMATION AND JOINT PREDICTION

The controlled experiments demonstrate two benefits of distributional correction. Auxiliary recording information reveals conditional changes invisible to native prediction inputs. Joint sampling then preserves associations lost by point predictions or independently combined marginals. Known binary and continuous distributions give exact targets against which to assess identification and fitting error.

D.1 SHIFTS WITH UNCHANGED MARGINALS

Generating law. Let X be uniform on $\{-1, 1\}$, $Z \sim \mathcal{N}(0, 1)$ independently, and $H = (X, Z)$. The source has $P_1(C = 1 \mid H) = 1/2$; deployments have $\pi_0(1 \mid H) = 1/2 + \kappa X$, for $\kappa \in \{-1/4, 1/4\}$. The stable response law is $P(Y = 1 \mid H, C) = 0.1 + 0.8C$ in the informative condition and $1/2$ in the unrelated-state control. The deployments share the same H, C and Y marginals but reverse the association between H and C. Their informative H-only probabilities are $1/2 + 0.8\kappa X$, with vector Brier risk 0.42 at the optimum, or 0.18 when the individual state is supplied. Indistinguishable inputs impose a direction-averaged bound of 0.50 (Appendix B.2). Joint (H, C) records distinguish the deployments.

Shared data and evaluation. Sixteen independent source samples each supply 512 labels, 1,024 source auxiliary records, 1,024 target auxiliary records per direction and 256 evaluation inputs. Balanced X pairs share Z; swapping states within pairs reverses the shift while keeping pooled counts fixed. Informative and unrelated-state conditions share feature and state streams. Exact expected vector Brier loss removes evaluation-label noise. Risk intervals pair directions within each source sample, and all model comparisons reuse these samples.

The exact ratio is $\Gamma(H, C) = 1 + 2\kappa X(2C - 1) \in [0.5, 1.5]$. Fitted probabilities use half-count smoothing within the two known X strata, supplying the relevant feature for estimating the correction. Appendix E.3 varies auxiliary-sample size. Native TFM prediction and its state-conditional components use all 512 labels and eight ensembles. Uniform and Context Direct use paired contexts of 205 examples, differing only in donor weights. Label access is matched; context length and inference cost differ. Source and deployment assembly share components. Pooled-state and most-probable-state controls test the need for conditional probabilities and averaging. Interfaces are specified in Appendix C.4.

Comparisons and inference. Two prespecified families use one-sided procedure-t bounds with Bonferroni correction: twelve gains over native prediction (three TFMs, two routes, two directions), and four gains comparing original TabICLv2 with SFLA and auxiliary source CRUMB with auxiliary deployment CRUMB in both directions. Each family uses a 0.005 useful-gain margin. Other intervals are descriptive, pointwise 95% intervals over the sixteen source samples.

D.1.1 NATIVE PREDICTION AND CONDITIONAL CORRECTION

Native inference receives C in labelled contexts and missing values in every inference row. Despite their different encodings, all three models return identical predictions across deployments, obeying the 0.50 average-risk bound. Context Direct and SFLA improve in both directions, and all twelve simultaneous lower bounds exceed 0.005 (Table D.1). SFLA approaches the attainable H-only risk of 0.42.

Updating the same components yields about 0.0804 for each model, with positive paired intervals. Pooled assembly stays near 0.50, selecting one state loses to averaging, and Context Direct improves on Uniform. The correction therefore goes beyond retaining the missing feature or resampling examples (Figure D.1).

D.1.2 MARGINAL CORRECTION AND REPEATED REGRESSION

DistPFN and its temperature variant postprocess fixed classifier outputs using deployment batches. An auxiliary global correction additionally uses recording-weighted source labels (Appendix E.2). Neither the batches nor a single class multiplier can recover the opposite conditional changes in this example. Their risks remain near 0.50 (Table D.2).



Figure D.1: Missing-state inputs and recording correction. (a) The tested APIs handle the missing column differently, but each receives identical inputs in the two deployments requiring opposite conditional predictions. (b) Direct and SFLA use unlabelled deployment records containing (H, C) while predicting from H alone. Evaluating possible states does not supply the individual's state. (c) Vector-Brier gains over native missing-state prediction, with paired pointwise 95% intervals after averaging directions within each of sixteen source samples. The unrelated-state control removes state–outcome dependence. The samples are reused from the earlier experiment, and context lengths differ across procedures.

Table D.1: Native prediction and recording correction over sixteen source samples. All methods predict without the individual's C. Original fits H alone; Native uses each released interface. (a) Mean vector Brier risk ↓ in both deployment directions; the unrelated-state control removes C's effect on the response. (b) Native-minus-corrected gains and simultaneous one-sided lower bounds for twelve prespecified comparisons. Every bound exceeds the useful-gain margin 0.005.

(a) Risks and controls

Model	Procedure	Informative state $\kappa = -.25$	Informative state $\kappa = +.25$	Unrelated state $\kappa = -.25$	Unrelated state $\kappa = +.25$
TabICLv2	Original (H only)	0.5045	0.5019	0.5025	0.5025
	Native	0.5045	0.5019	0.5026	0.5026
	Uniform contexts	0.5271	0.5368	0.5338	0.5338
	Context Direct	0.4633	0.4528	0.5311	0.5353
	Source assembly	0.5008	0.5026	0.5032	0.5032
	SFLA	0.4211	0.4212	0.5036	0.5037
	Pooled deployment assembly	0.5018	0.5000	0.5032	0.5032
	Most-probable state	0.5104	0.5124	0.5051	0.5053
TabDPT	Original (H only)	0.5020	0.5010	0.5010	0.5010
	Native	0.5095	0.5066	0.5005	0.5005
	Uniform contexts	0.5105	0.5084	0.5109	0.5109
	Context Direct	0.4427	0.4349	0.5131	0.5179
	Source assembly	0.5030	0.5006	0.5010	0.5010
	SFLA	0.4214	0.4214	0.5010	0.5012
	Pooled deployment assembly	0.5040	0.4981	0.5010	0.5010
	Most-probable state	0.4927	0.4984	0.5012	0.5017
TabPFN 3.5	Original (H only)	0.5019	0.5012	0.5011	0.5011
	Native	0.5015	0.4995	0.5009	0.5009
	Uniform contexts	0.5036	0.5080	0.5067	0.5067
	Context Direct	0.4368	0.4316	0.5067	0.5091
	Source assembly	0.5023	0.5014	0.5011	0.5011
	SFLA	0.4214	0.4215	0.5012	0.5014
	Pooled deployment assembly	0.5033	0.4989	0.5011	0.5011
	Most-probable state	0.4943	0.4985	0.5016	0.5021
	Attainable risk using H	.4200	.4200	.5000	.5000

(b) Prespecified correction gains

Model	Correction	Gain over native $\kappa = -.25$	Gain over native $\kappa = +.25$	Simultaneous lower bound $\kappa = -.25$	Simultaneous lower bound $\kappa = +.25$
TabICLv2	Context Direct	0.0412	0.0491	0.0321	0.0404
TabICLv2	SFLA	0.0834	0.0807	0.0806	0.0765
TabDPT	Context Direct	0.0668	0.0717	0.0574	0.0634
TabDPT	SFLA	0.0881	0.0852	0.0844	0.0825
TabPFN 3.5	Context Direct	0.0647	0.0679	0.0562	0.0598
TabPFN 3.5	SFLA	0.0801	0.0780	0.0780	0.0761

Our CRUMB paper reimplementation uses 64 random Fourier features, greedy batches of 50 and 205 distinct donors. Ordinary CRUMB matches visible features; auxiliary source and deployment variants match (H, C) with the same feature map, bandwidth, clusters and donor budget. The auxiliary update passes its simultaneous useful-gain tests, but gains over the original predictor are smaller: 0.0085 $[-0.0011, 0.0182]$ and 0.0104 $[0.0019, 0.0190]$. SFLA's gains over original TabICLv2 also pass the four-comparison test.

Repeated regression. The conditional mean specifies a Bernoulli distribution, so repeated regression (Boeken et al., 2023) gives a probabilistic comparator here. It reuses the logistic components and regresses their soft predictions on the 1,024 deployment auxiliary records onto $H = (X, Z)$, using ordinary least squares with an intercept. This primary estimator uses the same inference features as the other models. An X-only control averages pseudo-responses within $X = -1$ and $X = +1$, matching SFLA's state-probability model and discarding the independent noise feature Z. Outputs

Table D.2: Conditional correction under unchanged batch marginals. Sixteen independent procedures predict from H alone. (a) Mean Brier risk ↓ for additional comparators; global corrections marked ‡ reuse fixed classifier outputs. The full TFM controls are in Table D.1, and logistic means in Table D.3. (b) Bonferroni one-sided lower bounds for the four primary gains, with useful margin 0.005. (c) Comparator minus TabICLv2 SFLA with pointwise 95% paired intervals. The control with a state unrelated to the response retains the small losses from estimated assembly.

(a) Expected Brier risks

	Informative state		Unrelated state	
Procedure	$\kappa = -.25$	$\kappa = +.25$	$\kappa = -.25$	$\kappa = +.25$
Original TabICLv2	0.5045	0.5019	0.5025	0.5025
TabICLv2 SFLA	**0.4211**	**0.4212**	0.5036	0.5037
DistPFN‡	0.5048	0.5022	0.5027	0.5027
DistPFN-T‡	0.5037	0.5012	**0.5020**	**0.5020**
Auxiliary global correction‡	0.5040	0.5018	0.5028	0.5027
Visible-feature CRUMB	0.5108	0.5132	0.5082	0.5082
Auxiliary CRUMB: source	0.6529	0.5983	0.5096	0.5096
Auxiliary CRUMB: deployment	0.4959	0.4915	0.5097	0.5087
Attainable risk using H	.4200	.4200	.5000	.5000

(b) Prespecified gains and simultaneous lower bounds

Contrast	κ	Gain [95% interval]	Simultaneous lower bound
Original → SFLA	$-.25$	0.0833 [0.0814,0.0853]	0.0811
Original → SFLA	$+.25$	0.0807 [0.0777,0.0836]	0.0772
Auxiliary CRUMB: source → deployment	$-.25$	0.1570 [0.1198,0.1942]	0.1135
Auxiliary CRUMB: source → deployment	$+.25$	0.1068 [0.0751,0.1386]	0.0697

(c) Additional paired comparisons

Comparator	Outcome law	$\kappa < 0$	$\kappa > 0$
DistPFN-T	Informative	$+0.0826$ $[+0.0806, +0.0846]$	$+0.0799$ $[+0.0772, +0.0827]$
DistPFN-T	Null	-0.00155 $[-0.00222, -0.00088]$	-0.00163 $[-0.00251, -0.000756]$
Auxiliary global correction	Informative	$+0.0828$ $[+0.0805, +0.0852]$	$+0.0805$ $[+0.0770, +0.0840]$
Auxiliary global correction	Null	-0.000796 $[-0.00168, +0.000091]$	-0.000917 $[-0.00161, -0.000225]$
Logistic SFLA	Informative	$+0.00078$ $[+0.000262, +0.00130]$	$+0.00036$ $[+0.000026, +0.000693]$
Logistic SFLA	Null	-0.000667 $[-0.00162, +0.000287]$	-0.000528 $[-0.00155, +0.000493]$

are clipped to $[0, 1]$; clipping changes no evaluated prediction. These outer regressors differ from the thin-plate splines in the authors' continuous experiments.

Both repeated regression and SFLA recover the correction (Table D.3). Primary repeated regression improves the original by 0.0811 $[0.0797, 0.0825]$. SFLA improves on it by 0.00028 $[0.00012, 0.00043]$, while the X-only control improves on SFLA by 0.00017 $[0.00005, 0.00029]$. These small differences depend on the outer estimator. Replacing deployment auxiliary records with source records gives no correction. Removing SFLA's half-count smoothing changes informative risk by less than 0.00001. With an unrelated state, primary repeated regression loses 0.00038 $[0.00006, 0.00071]$ to the original. These paired intervals measure risk differences on inspected samples; they do not assess calibration. The continuous comparison is in Appendix F.1.1.

When no conditional correction is needed. With $Y \perp C \mid H$, Context Direct increases risk over native prediction by about 0.031, 0.015 and 0.007 for TabICLv2, TabDPT and TabPFN. Much of the loss is already present in uniform resampling. TabDPT also loses from targeting relative to that control. SFLA's smaller losses (0.0003–0.0011) have positive paired loss intervals for all three models. In the H-only comparison, DistPFN-T beats SFLA by about 0.0016 in both directions (Table D.2). These secondary comparisons have no equivalence margin. The component calculation in Appendix E.3 explains why even exact mixture probabilities can leave a loss.

Table D.3: Repeated regression and assembly with the same fitted logistic components. Mean expected vector-Brier risk ↓, averaging the two deployment directions within each of sixteen source streams. Predictive and unrelated refer to the response's dependence on C given H. Bold marks the lowest displayed means.

Procedure	Predictive C	Unrelated C
Original logistic	0.5031	0.5027
Weighted Direct	0.4228	0.5034
Source assembly	0.5022	0.5027
SFLA	0.4218	0.5030
Repeated regression	0.4220	0.5030
Repeated regression (X only)	**0.4216**	**0.5022**
Repeated regression (source auxiliary sample)	0.5024	0.5027

D.2 Continuous prediction with indistinguishable native inputs

Direct and SFLA improve continuous prediction when native inputs and unconditional marginals cannot distinguish the deployments (Table D.4). Both flow implementations improve using probabilities estimated from auxiliary records. Competitive Gaussian controls show that the same correction also works with simpler models.

Let H be uniform on $\{-1, 1\}$, with source $P_1(C = 1 \mid H) = 1/2$ and deployment probabilities $\pi_r(1 \mid h) = 1/2 + rh/4$, $r \in \{-1, 1\}$. For $s_c = 2c - 1$,

$$Y \mid H, C = c \sim N\left(s_c(0.6, -0.6)^\top, 0.64 \begin{pmatrix} 1 & 0.4s_c \\ 0.4s_c & 1 \end{pmatrix}\right).$$

Both deployments preserve $P(H)$ and the entire unconditional response law. Their conditional coordinate marginals can change, unlike the separate dependence-only experiment in Appendix D.3. The null retains the source H, Y arrays but makes recorded C independent of the latent state generating Y.

Four independent fitting streams each provide 4,096 training labels, 2,048 separate source validation labels and independent outcome-free source and deployment auxiliary samples of 2,048 observations each. Half-count estimates of the state probabilities supply the correction. Within each flow implementation, source and weighted training share the input layout and initialization. Both run for 32,768 updates, retaining checkpoints at 256, 1,024, 8,192 and 32,768 updates. Original prediction uses source-only fitting and stopping. Direct weights both. The two assemblies reuse the same source-selected components and generated values, changing only the mixture probabilities. Gaussian and GMM2 controls condition on H and, for components, on C; weighted GMM training uses importance resampling.

In a secondary comparison computed after evaluation, weighted training gains 0.0294 $[0.0273, 0.0315]$ and 0.0288 $[0.0253, 0.0323]$ with weighted stopping on both trajectories. The exact source-to-deployment opportunity is 0.031733. A fitted original's additional estimation error allows its improvement to exceed this value without outperforming the deployment oracle. Replacing estimated deployment probabilities by oracle probabilities reduces SFLA risk by about 0.0002 for either flow. The null intervals retain uncertainty about harm, including losses larger than 0.005 for ordinary-FM assembly.

D.3 Joint prediction and completed-table quantities

Recording changes dependence while preserving every scalar marginal. Four responses form two independent pairs conditional on visible group G, recording state C and query A. For either pair (j, k),

$$q_c(y_j, y_k \mid a) = 2\phi(y_j)\phi(y_k)\sigma\{(2c - 1)k_a y_j y_k\}, \qquad (k_0, k_1) = (0.75, 1.25).$$

Gaussian symmetry and $\sigma(u)+\sigma(-u) = 1$ give standard normal marginals in both states. Conditional means are zero, but product moments change sign: exact mean predictions therefore provide a useful point-output control for completed-table products and slopes.

Table D.4: Continuous recording correction. (a) Mean energy/$\sqrt{2}$ over four fitting streams; bold marks the lowest displayed mean within each row. (b) Paired gains with descriptive 95% t intervals across those streams. Directions and two simulation banks are averaged within each stream. Positive gains favor correction.

(a) Predictive risk

Condition	Model	Original	Direct	Source assembly	SFLA
Informative	Ordinary FM	0.6248	0.5933	0.6239	**0.5926**
	Matched CFMI	0.6255	0.5942	0.6235	**0.5921**
	Gaussian	0.6249	0.5935	0.6232	**0.5917**
	GMM2	0.6234	0.5920	0.6233	**0.5917**
State-independent	Ordinary FM	0.6248	**0.6237**	0.6252	0.6254
	Matched CFMI	0.6255	**0.6244**	0.6253	0.6253
	Gaussian	0.6249	0.6250	**0.6247**	0.6248
	GMM2	0.6234	0.6241	**0.6231**	0.6233

(b) Correction gains

Comparison	Ordinary FM	Matched CFMI
Informative recording state		
Original − Direct	$+0.0315$ $[+0.0278, +0.0352]$	$+0.0313$ $[+0.0268, +0.0358]$
Original − SFLA	$+0.0322$ $[+0.0284, +0.0360]$	$+0.0335$ $[+0.0310, +0.0360]$
Source assembly − SFLA	$+0.0313$ $[+0.0305, +0.0320]$	$+0.0314$ $[+0.0309, +0.0320]$
State independent of the response		
Original − Direct	$+0.0012$ $[-0.0018, +0.0041]$	$+0.0012$ $[-0.0003, +0.0027]$
Original − SFLA	-0.0006 $[-0.0081, +0.0070]$	$+0.0003$ $[-0.0012, +0.0017]$
Source assembly − SFLA	-0.0002 $[-0.0012, +0.0008]$	$+0.0000$ $[-0.0006, +0.0006]$

Generating law and data. Visible group and state have uniform joint mass before recording. State-specific recording probabilities are $0.75, 0.25$. Source probabilities of query $A = 1$ given (G, C) are $\left[\begin{smallmatrix} .25 & .45 \\ .40 & .60 \end{smallmatrix}\right]$ and deployment probabilities are $\left[\begin{smallmatrix} .65 & .45 \\ .50 & .30 \end{smallmatrix}\right]$, with rows indexing group and columns state. Sixteen independent procedures each use 4,096 labelled source training rows, 2,048 separate source-selection rows and 512 source rows for numerical checks. Evaluation completes 128 missing blocks alongside 128 recorded blocks with 64 fresh predictive draws per missing block. We fix all forecasts before using deployment outcomes for evaluation.

Each fitted library returns the component given (G, C, A), a source mixture given (G, A), and a deployment mixture at the same inference input. A single sampled state governs the whole vector. Mixture probabilities and source importance weights are exact design quantities, including the query policy change. This isolates component fitting and mixture use from estimation of the recording mechanism.

Predictors and selection. Training follows Table C.3. DT-FM uses 256 midpoint steps, CFMI uses 100/400 Euler steps and DiffPuter retains its stochastic sampler. CFMI's terminal predictions at both resolutions separate solver and checkpoint effects. Gaussian, GMM4, GMM8 and copula models fit the eight (G, C, A) cells. GMM8 meets its convergence and effective-count requirements in nine selected procedures; the other families contribute sixteen. Selection uses weighted source energy with 256 draws per component. Evaluation draws are independent of selection.

Association targets and inference. Completed tables are assessed through products $Y_j Y_k$ and ordinary least-squares slopes with an intercept. Each slope is the empirical covariance divided by the predictor variance, averaged across completed draws. The reference combines the observed recorded rows with true conditional moments at each missing row's realized state: zero means, unit variances and product $(2C - 1)\kappa_A$. This ratio of expected moments differs from the expected finite-table slope. It also uses the actual missing-row states, whereas the deployed predictor sees only (G, A).

The primary tests evaluate selected DT-FM with C supplied. Product and slope biases for both pairs must have 90% intervals within ± 0.03; one-sided upper bounds on absolute-bias differences

from independent exact marginals must be below -0.02. All four checks pass (Table D.6). Other comparisons use descriptive pointwise 95% intervals. The later Gaussian check provides an oracle matched to H-only inference (Appendix D.3.1).

Dependence and mixture controls. For each predictive sample set, taking the product of its coordinate-wise empirical distributions preserves every marginal exactly. Its expected product is the product of its coordinate means and needs no additional sampling. Selected DT-FM's primary product bias changes from $+0.00265$ to -0.07768 under this transformation. Table D.6 reports both response pairs across the sampler families. Gaussian and GMM4 recover the product too: covariance suffices for this particular functional.

Point imputers and joint TabICLv2. Matched ReMasker, PMAE and HyperImpute fits use the same sixteen source assignments. The masked autoencoders retain their 600-epoch point-output configurations. PMAE hides one eligible target per source row. HyperImpute uses 20% single-target amputation. Configurations are fixed before evaluation.

TabICLv2 uses eight regression ensembles and scalar conditional sampling in the fixed order $(0, 1, 2, 3)$, feeding each generated prefix into the next prediction. State-conditional predictions also receive C. Assembly draws one state for the whole vector. We use two independence controls: generating each response without earlier sampled values, and independently recombining the completed samples. Only the second preserves the empirical marginals exactly. Context Direct and Uniform use resampled contexts of 4,096 examples without trained heads.

Point predictions approach the squared-loss optimum but have second moments below 0.006, so completed values lose variability. Joint TabICLv2 recovers spread and reduces primary product bias from -0.1544 under source assembly to $+0.0039$ under deployment assembly. Its paired energy gain is 0.00182 $[0.00106, 0.00257]$. Context Direct reduces association bias relative to Uniform without a resolved energy gain. SFLA improves both quantities relative to Direct (Table D.5). These methods share the sixteen assignments. The primary DT-FM tests remain separate.

Table D.5: Joint TabICLv2 on the four-response law. (a) Energy/2, marginal CRPS and primary product bias; point rows return deterministic predictions. (b) First-minus-second differences with pointwise 95% intervals over the same sixteen assignments. Direct and Uniform use contexts of 4,096 examples; joint prediction follows coordinate order $(0, 1, 2, 3)$.

(a) Prediction accuracy

Predictor	Energy ↓	Marginal CRPS ↓	Product bias
Original joint TFM	0.6680	0.5654	-0.1576
Uniform-context joint TFM	0.6684	0.5656	-0.1536
Context Direct joint TFM	0.6693	0.5679	-0.0439
Source assembly	0.6680	0.5652	-0.1544
Deployment assembly	0.6662	0.5654	+0.0039
Independent scalar assembly	0.6668	0.5656	-0.0780
Native TFM point	0.9467	0.8009	-0.0794
Assembled TFM point	0.9470	0.8011	-0.0791

(b) Paired gains

Paired difference	Energy difference [95% interval]	Absolute product-bias difference
Source minus deployment	+0.0018 [+0.0011, +0.0026]	+0.1397 [+0.1274, +0.1520]
Original minus deployment	+0.0018 [+0.0011, +0.0024]	+0.1429 [+0.1299, +0.1559]
Uniform minus Direct	-0.0008 [-0.0024, +0.0007]	+0.1097 [+0.1029, +0.1166]
Direct minus deployment	+0.0031 [+0.0019, +0.0043]	+0.0292 [+0.0164, +0.0420]

Write $\kappa_a = \mathbb{E}[Y_0 Y_1 \mid C = 1, A = a] > 0$ and $\pi_r(g, a) = P_r(C = 1 \mid G = g, A = a)$. Then

$$\mathbb{E}_{Q_r}[Y_0 Y_1 \mid g, a] = (2\pi_r(g, a) - 1)\kappa_a, \qquad \mathbb{E}_{Q_1 - Q_0}[Y_0 Y_1 \mid g, a] = 2\{\pi_1(g, a) - \pi_0(g, a)\}\kappa_a.$$

Here $\kappa_0 = 0.29341$ and $\kappa_1 = 0.39430$. Source state-one probabilities are below $1/2$ in every cell, while deployment probabilities exceed $1/2$. Source assembly therefore replaces positive target dependence by negative dependence, whereas independent marginals replace it by zero. The completed-table effect is scaled by its missing fraction, here $1/2$. This explains why the source-mixture oracle can be worse than independent exact marginals.

Table D.6: Joint prediction at supplied state. (a) Product and slope biases satisfy the ± 0.03 equivalence tolerance and the required paired absolute-bias reduction. (b) Descriptive product/slope biases target zero, coverage targets 90%, and energy is divided by two. Native denotes terminal flows, training-loss-selected DiffPuter or fitted statistical models; selected uses source energy of the H-only mixture. Selected GMM8 has nine eligible fits; its native row and all other rows have sixteen.

(a) Prespecified DT-FM decision at supplied state

Pair	Quantity	Mean bias	90% interval	Absolute-bias difference one-sided 95% upper bound
$(0, 1)$	Product	+0.0027	$[-0.0026, +0.0079]$	-0.0610
$(0, 1)$	Slope	+0.0028	$[-0.0024, +0.0080]$	-0.0624
$(2, 3)$	Product	-0.0051	$[-0.0120, +0.0018]$	-0.0625
$(2, 3)$	Slope	-0.0048	$[-0.0113, +0.0017]$	-0.0650

(b) Fitted-model functionals and calibration

Model	Output	n	Energy ↓	Coverage (90%)	Product bias		Slope bias	
					Pair 1	Pair 2	Pair 1	Pair 2
DT-FM	Native	16	0.6676	86.9%	+0.0022	-0.0016	+0.0016	-0.0016
DT-FM	Selected	16	0.6679	87.0%	+0.0027	-0.0051	+0.0028	-0.0048
CFMI	Native	16	0.6665	86.5%	-0.0004	-0.0074	-0.0008	-0.0079
CFMI	Terminal, 400 steps	16	0.6665	86.8%	+0.0002	-0.0067	-0.0001	-0.0071
CFMI	Selected	16	0.6675	86.9%	-0.0337	-0.0441	-0.0355	-0.0456
DiffPuter	Native	16	0.6874	85.6%	-0.0035	+0.0133	-0.0059	+0.0060
DiffPuter	Selected	16	0.6790	86.3%	+0.0130	+0.0120	+0.0089	+0.0078
Gaussian	Native = selected	16	0.6677	87.1%	+0.0038	-0.0024	+0.0034	-0.0028
GMM4	Native = selected	16	0.6677	87.2%	+0.0034	-0.0023	+0.0031	-0.0028
GMM8	Native	16	0.6686	86.9%	+0.0044	-0.0024	+0.0040	-0.0031
GMM8	Selected	9	0.6646	87.2%	+0.0029	+0.0030	+0.0032	+0.0025
Copula	Native = selected	16	0.6677	86.9%	-0.0052	-0.0101	-0.0057	-0.0106

Checkpoint choice. Native CFMI has primary/replication product biases $-0.00041/-0.00736$, compared with $-0.03365/-0.04413$ after selection. The same terminal at 400 steps gives $+0.00025/-0.00668$. The difference therefore persists after matching the solver. CFMI selects an earlier checkpoint in eight procedures and its terminal in eight. DT-FM selects its terminal in thirteen. Selection minimizes energy for predictions that omit C. The product check instead evaluates a particular association conditional on C. A checkpoint can therefore improve the selection score while worsening this association.

Table D.7: Selected H-only assemblies: mean energy and paired source-minus- deployment gain over the available independent procedures. GMM8 is evaluated on the nine procedures that satisfy its eligibility checks; the other families use sixteen.

Model	n	Source risk	Deployment risk	Gain	Paired 95% interval
DT-FM	16	0.6704	0.6686	+0.0018	$[-0.0005, +0.0041]$
CFMI	16	0.6709	0.6681	+0.0028	$[+0.0004, +0.0053]$
DiffPuter	16	0.6766	0.6781	-0.0015	$[-0.0052, +0.0022]$
Gaussian	16	0.6698	0.6683	+0.0015	$[-0.0015, +0.0046]$
GMM4	16	0.6697	0.6685	+0.0012	$[-0.0012, +0.0037]$
GMM8	9	0.6651	0.6652	-0.0001	$[-0.0033, +0.0031]$
Copula	16	0.6697	0.6683	+0.0014	$[-0.0016, +0.0045]$

Most energy-gain intervals include zero despite the correction in product bias. The experiment tests product and slope accuracy. Downstream inferential coverage would require a separate evaluation.

D.3.1 A GAUSSIAN DEPENDENCE-ONLY CHECK

A second Gaussian law provides a check against the oracle available from H alone. For independent standard Gaussian Z, E,

$$Y_1 = Z, \qquad Y_2 = 0.9(2C-1)Z + \sqrt{1-0.9^2}\,E.$$

Both scalar marginals are standard Gaussian in either state. Source and deployment state-one probabilities are 0.8 and 0.2. The input H is independent of response and state. Each of three DT-FM fits uses 2,048 fitting rows, 512 separate validation rows and 2,048 evaluation contexts, with the application training configuration (32,768 updates, 512 midpoint steps). Risk averages 64 exact-law outcome draws per evaluation context. Each component library serves both source assembly and SFLA (Table D.8). A positive mean joint-score gain accompanies a small negative mean marginal-score gain. Population marginal opportunity is exactly zero, so the fitted marginal difference reflects estimation and sampling error.

Table D.8: Gains from updating a fitted library with unchanged population marginals, mean (fit SD) over three seeds. Positive favors updating. Each score is reported in its own units.

Score	Source-assembly minus SFLA risk
Joint energy	0.0250 (0.0115)
Mean marginal CRPS	-0.0043 (0.0089)

The oracle matched to H-only inference draws a fresh state from the deployment mixture for each sample. Selected excess energy above this oracle is 0.0136 (fit SD 0.0087), versus terminal excess 0.3184 (SD 0.0142). The fitted mixture-update gain, 0.0250 (SD 0.0115), is compared with an exact-component opportunity of about 0.0300. The additional value of observing each individual's state is 0.0138 (SD 0.0004), close to the separately estimated projection term 0.0133 in Proposition 1. This separates the information unavailable to the predictor from its remaining estimation error.

E FITTED RECORDING CORRECTION

The fitted-error analysis explains why updating the same components can improve prediction in one population and harm it in another. These experiments measure how the gain depends on component accuracy, estimated mixture probabilities and departures from conditional stability.

E.1 SHIFT RESPONSE WITH FIXED COMPONENTS

A four-dimensional Gaussian response depends on two context features. State changes its conditional mean and reverses correlations within two response pairs. In the null condition, both state-conditional laws are identical. DT-FM and four-component Gaussian mixtures are fitted at source state-one probability 0.7. Three independent fits per law each serve seven shifts $x \in \{0, \pm 0.05, \pm 0.1, \pm 0.2\}$ with visible contexts fixed. Each curve $G(x) = ax + bx^2$ keeps its fitted components fixed across shifts, so differences between the three curves reflect fitting variation. Risks use common coordinates and fair pairwise scoring, with unchanged predictions at zero shift. The classifier study below also tests updates when the state carries no additional outcome information.

When the recording state predicts the four-dimensional response, mean DT-FM gains at $x = -0.2$ and $x = 0.2$ are 0.004119 and 0.004522. The oracle quadratic opportunity at either shift is 0.004533. The fitted response coefficients in $G(x) = ax + bx^2$ are

$$a = 0.001006 \quad [-0.010054, 0.012066], \qquad b = 0.108016 \quad [0.095753, 0.120279],$$

with true quadratic coefficient $d = 0.113331$. The quadratic coefficient is about 95.3% of d across three fits. Null gains at the extremes are 0.000062 and -0.000152.

E.2 MIXTURE ESTIMATION FOR FROZEN CLASSIFIERS

Figure 2 compares source-predicted gains with independent evaluation. Here changing the state proportion also changes visible contexts: draw $C \sim \text{Bernoulli}(q)$ and independent standard Gaussian

features, then shift H_1 by $0.4(2C-1)$. With logistic function σ,

$$P(Y=1 \mid H, C) = \sigma(0.7H_1 - 0.6H_2 + 0.45H_3H_4 + (2C-1)\{1.4 + 0.5\tanh H_2\}) .$$

The control with a state unrelated to the response drops the final state-dependent term, so $Y \perp C \mid H$ while state proportions can still change. The exact contextual state probability is $\sigma\{\mathrm{logit}(q)+0.8H_1\}$. Source state-one probability is 0.8, and deployments use $0.2, 0.4, 0.6, 0.8, 0.9$.

Sixteen independent procedures each use 512 fitting labels, 1,024 source records for the state-probability model, 1,024 separate source labels to predict gain and 256 outcome-free target states per composition. The target frequency is unsmoothed. A separate 1,024-query sample from prior 0.5 is reweighted to each deployment. Components, probability estimates and evaluation queries are shared across compositions within a procedure; forecasts are fixed before evaluation outcomes are drawn.

At target prior 0.2, source losses predict gain 0.16550, compared with observed gain 0.17324 $[0.157, 0.189]$. Their paired difference is 0.00774 $[-0.00728, 0.02276]$, with prediction RMSE 0.02837. The known/estimated-proportion comparisons use Holm adjustment for two tests and have no equivalence margin. Without shift, updating loses 0.000697 $[0.000119, 0.001275]$. With an unrelated state its mean gain is -0.002604 $[-0.005410, 0.000203]$, with the interval including zero. After a 0.01167 loss from fitting the source mixture, the original-relative SFLA gain is 0.16156 $[0.14795, 0.17518]$.

Logistic regression also benefits: its Direct and SFLA risks are about 0.3801 and 0.3797, comparable to TabICLv2 SFLA at 0.3801. A simple model captures the conditional change in this example. The class-prior controls below establish how much of the gain comes from a global probability adjustment.

Class-prior correction. DistPFN (Lee, 2026) adjusts a classifier's probabilities using the class frequencies in its source-fitting labels and an unlabelled deployment batch. Write these frequencies as s_y and the mean batch prediction as m_y. Its prediction is

$$\widetilde{p}(y \mid h) = \frac{p(y \mid h)\, a_y/(s_y + 10^{-8})}{\sum_k p(k \mid h)\, a_k/(s_k + 10^{-8})}.$$

The standard variant sets $a = m$. The temperature variant uses $a = \mathrm{softmax}(m/\tau)$ with $\tau = -\sum_y m_y \log(\max\{s_y, 10^{-12}\})$, following the author's probability-based rule. Both reuse the same fitted classifier.

The auxiliary-prior control instead uses recording-weighted source labels, $a_y = \sum_i \widehat{\Gamma}_i \mathbf{1}\{Y_i = y\}/\sum_i \widehat{\Gamma}_i$. It receives the additional recording information available to Direct and SFLA. All three rules postprocess probabilities without deployment responses or individual evaluation states.

Two additional 16-procedure comparisons apply the class-prior corrections to actual deployment batches. The estimated-mixture comparison uses 512 fitting labels, 512 selection labels, 1,024 source auxiliary rows and 2,048 queries at each prior 0.2 or 0.8. Context Direct uses 256 examples per context, so its comparison with the 512-label original also changes context size. The known-prior comparison instead reuses the prospective-response libraries at their actual 1,024-query prior-0.5 batch, under the informative and target-null laws. No Direct fit exists at that composition. These are separate allocations from the reweighted-reference experiment above.

At target proportion 0.2, auxiliary class-prior correction recovers 98.02% of SFLA's original-relative gain, leaving an additional 0.00349 for SFLA. At 0.5 their ordering reverses slightly, with an interval including zero. Global correction therefore captures most of this law's benefit, unlike the unchanged-marginals experiment in Appendix D. With an unrelated state, both DistPFN variants and SFLA have resolved losses. Changing state composition can change H and class frequencies even when $Y \mid H$ is unchanged, so label-shift assumptions need not hold. Spurious fitted component differences can separately harm SFLA.

E.3 Probability estimation and component error

Table E.2 holds sixteen libraries fixed and averages over new auxiliary counts. Panel (a) includes half-count bias and variance with $m/2$ observations per X stratum (Equation (B.4)). Panel (b) integrates binomial target frequencies through the fitted state-probability model.

Table E.1: Class-prior and conditional-assembly controls: vector Brier ↓. The first two columns use actual deployment batches and the estimated-mixture libraries; the last two use the separate known-prior reference comparison. A dash denotes an unavailable procedure.

Procedure	$q = .2$, shift	$q = .8$, identity	$q = .5$, shift	$q = .5$, unrelated state
TabICLv2: original	0.5518	**0.3642**	0.4520	**0.4228**
TabICLv2: source assembly	0.5529	0.3645	0.4580	0.4239
TabICLv2: SFLA	**0.3756**	0.3645	0.4118	0.4245
TabICLv2: DistPFN	0.4905	0.3654	0.4362	0.4250
TabICLv2: DistPFN-T	0.4653	0.3649	0.4233	0.4248
TabICLv2: auxiliary class prior	0.3791	**0.3642**	**0.4110**	0.4248
TabICLv2: Context Direct	0.3973	0.3711	—	—
Logistic: Direct	0.3788	0.3654	—	—
Logistic: SFLA	0.3784	0.3659	0.4118	0.4291
Tree: Direct	0.5336	0.4289	—	—
Tree: SFLA	0.4046	0.3995	0.4374	0.4910

Table E.2: Auxiliary-sample effects for two fixed-library experiments. (a) Risks include half-count bias and variance. (b) Gains average over binomial target-frequency estimates. The same fitted models are reused at every auxiliary sample size.

(a) Unchanged marginals: expected Brier risk ↓

	Informative state		Unrelated state	
Auxiliary rows	$\kappa < 0$	$\kappa > 0$	$\kappa < 0$	$\kappa > 0$
128	0.4245	0.4246	0.5036	0.5037
512	0.4217	0.4218	0.5036	0.5037
1024	0.4212	0.4213	0.5036	0.5037
4096	0.4208	0.4209	0.5036	0.5037
Exact probabilities	0.4207	0.4208	0.5036	0.5037

(b) Prospective law: mixture-update gain ↑

Auxiliary states	Informative, $q = .2$	Informative, $q = .8$	Null, $q = .2$
64	+0.1718	-0.001611	-0.002709
256	+0.1730	-0.000401	-0.002691
1024	+0.1733	-0.000100	-0.002687
4096	+0.1733	-0.000025	-0.002686

At 256 states in the prospective law, estimating the shifted prior costs 0.000381 $[0.000301, 0.000460]$ relative to the exact prior. With unchanged prior it incurs 0.000401 $[0.000355, 0.000447]$ unnecessary loss. With an unrelated state, the update still harms prediction when state probabilities are known exactly. More auxiliary data cannot remove the spurious differences between the fitted state-conditional predictions. The near-$1/m$ decrease concerns probability-estimation error in this experiment.

Component error in the unchanged-marginals law. For the TabICLv2 components of Appendix D.1, the informative-condition Brier component-error term is about 0.0007–0.0008, state-probability error 0.00049, and their interaction is negative. Exact-gate updating averages 0.079863 across the two directions, close to the population opportunity 0.08. Under the null, the fitted component contrast creates error even with exact probabilities, as the next calculation shows.

Let $m = (\widehat{q}_0 + \widehat{q}_1)/2$, $\widehat{d} = \widehat{q}_1 - \widehat{q}_0$, and $u = (1/2, 1/2)$. Under the null, exact gates return $m \pm \frac{1}{4} X\widehat{d}$. Their average excess Brier risk is

$$\mathbb{E}_H \|m - u\|^2 + \frac{1}{16}\mathbb{E}_H \|\widehat{d}\|^2 = 0.003172 + 0.000499.$$

The signed cross term cancels between opposite shifts. The average component error already exceeds the original predictor's excess 0.002546. Here, suppressing the fitted state contrast removes part of the loss, but the source mixture can still be worse than the original predictor. The two remain separate fallback choices.

E.4 Sensitivity to unobserved outcome change

Context Direct and SFLA improve over the original at every tested departure in two specified families, while their ordering can reverse. The test reuses the sixteen estimated-mixture procedures from Appendix E.2, holding the source law, deployment (H, C), forecasts, estimated mixture probabilities and selection decisions fixed while varying the unavailable deployment outcomes. If $q_c(h) = P_1(Y = 1 \mid h, c)$, set

$$q_{c,\eta}(h) = q_c(h) + \eta d_c \min\{q_c(h), 1 - q_c(h)\}, \qquad \eta \in \{0, \pm 0.1, \pm 0.25, \pm 0.5\}.$$

The two prespecified directions are a common outcome shift $d_c = 1$ and a state contrast $d_c = 2c - 1$. The resulting conditional density ratio lies in $[1-|\eta|, 1+|\eta|]$. These populations are indistinguishable from the learner's observed data, although recording can now depend on the outcome given (H, C).

For source/deployment state proportions $0.8/0.2$, integrate the two states and binary outcomes at each saved query. Table E.3 reports expected summed two-class Brier gains, with paired pointwise t intervals across sixteen procedures. This removes outcome sampling noise while retaining query and fitting variability. All departure levels reuse the same fitted procedures.

Table E.3: Both corrections retain gains under the tested departures; the Context Direct–SFLA ordering reverses. Gains are relative to the original frozen TabICLv2; positive favors the correction. Context Direct targets demonstrations; SFLA updates the mixture probabilities. Intervals are paired 95% intervals over sixteen procedures.

Departure η	Context Direct gain	SFLA gain	SFLA gain minus Context Direct gain
-0.5	0.2363 $[0.2202, 0.2524]$	0.2976 $[0.2817, 0.3135]$	0.0613 $[0.0481, 0.0745]$
$+0.0$	0.1596 $[0.1468, 0.1723]$	0.1824 $[0.1674, 0.1973]$	0.0228 $[0.0150, 0.0307]$
$+0.5$	0.0828 $[0.0721, 0.0935]$	0.0671 $[0.0513, 0.0830]$	-0.0157 $[-0.0234, -0.0080]$

SFLA gains 0.1824 at the assumed bridge and 0.0671 at $\eta = 0.5$, where Context Direct performs better. SFLA's excess over the perturbed Bayes law increases from 0.0165 to 0.0482. The state-contrast departure has a different direction of effect: SFLA gains 0.1125 at $\eta = -0.5$ and 0.2522 at $\eta = 0.5$. Both methods retain positive original-relative gains at every tested departure level in these two directions, while their ranking and distance from the correct law depend on the departure.

The sensitivity identity in Appendix B.9 bounds how gain changes from its nominal value. For original and corrected class-one probabilities $b(h), f(h)$, writing $\delta_\eta(h) = \sum_c \pi_0(c \mid h)\{q_{c,\eta}(h) - q_c(h)\}$ gives

$$G_\eta - G_0 = 4E[\delta_\eta(H)\{f(H) - b(H)\}].$$

The affine response follows from the identity. The experiments estimate its slope and show where the ranking reverses under the specified unobserved outcome changes. Because every case has the same observed data, source-based selection and a doubly robust objective also face this identification limit.

F Prediction on real datasets

The five-application comparison measures predictive accuracy and isolates the contribution of recording correction. Learning and selection experiments then examine how to obtain a useful final predictor. Completion benchmarks compare samplers with native imputers under both complete and selectively recorded training responses, followed by Adult classification and clinical laboratory prediction. Estimators and scoring rules follow Appendix C.

F.1 Recording correction across five applications

SFLA improves native TabPFN prediction on Household and Metro, and updating fixed components lowers mean risk for all displayed model families on both tasks. The comparisons below distinguish

Table F.1: Application designs and controls. Gas and Appliances grant auxiliary access to a response-derived state without its numerical outcomes. The other applications withhold an existing covariate from inference.

Application	Responses; recording information C	Control	Evaluation groups
Gas Turbine	CO, NO_x; product of residualized responses	State unrelated to response given H	141/143 time groups
Appliances	Temperature T2, humidity RH2; same state construction	State unrelated to response given H	25 days
Beijing	$PM_{2.5}$, PM_{10}; station, rates $.3/.7$	Equal station rates	60 weeks
Household	Active, reactive power; submeter-3 activity, rates $.3/.7$	Equal activity-state rates	80 weeks
Metro	Traffic volume; Saturday status, rates $.7/.3$	Equal day-state rates	99 weeks

these gains from context resampling, component fitting and native missing-feature handling. The other applications show how gains vary with recording design and fitted predictors.

Common design. Recording is imposed on measured outcomes, retaining the hidden responses for evaluation. Recording information C is available during estimation and omitted from individual inference inputs. Table F.1 defines the states and controls. Pretrained and statistical procedures use 2,048 fitting and 512 scaling labels. Flows add 256 validation labels (Appendix F.3).

Original pretrained predictors use H alone. Native TabDPT and TabPFN retain source C and use mean filling and missingness indicators, respectively, with eight ensembles and no deployment auxiliary records. Uniform and Context Direct average two contexts of 2,048 examples, differing only in donor weights. Source assembly and SFLA share component predictions and change their mixture probabilities. API details are in Appendix C.4.

We evaluate fair CRPS with 128 draws on source-fitted scales, averaging responses within each case. Paired, pointwise 95% temporal-group intervals condition on fitted predictors, estimated weights and sampled forecasts. The comparisons reuse the same allocations and provide no independent replication. Household uses Tuesday–Saturday hourly records, with voltage, submeters 1 and 2 and cyclic time as inputs, excluding global intensity, submeter 3 and both responses. Metro uses Tuesday–Saturday hourly traffic, weather, hour and season, withholding Saturday status (Hogue, 2019). Weeks separate fitting, auxiliary and evaluation data.

F.1.1 Predictive performance against existing methods

Table 3 reports absolute risks. The comparisons below test native prediction, context matching and repeated regression using the same application allocations.

Native prediction. SFLA improves on both native and H-only predictions for TabDPT and TabPFN on Household (Table F.2). TabPFN's mixture gain comes from active power (.0352); its reactive-power change is $-.0007$ $[-.0016, .0002]$. Context Direct also improves on native TabPFN by .0108 $[.0076, .0140]$.

Native prediction already improves both Metro originals. SFLA adds a clear gain for TabPFN and an uncertain one for TabDPT. TabPFN targeting improves uniform contexts by .0161 $[.0126, .0196]$, without a clear gain over native prediction. Native TabPFN outperforms SFLA on Beijing in both recording conditions and on Appliances at the primary score. The Appliances sign persists with 32/64 draws, though those intervals include zero. On Gas, TabDPT improves on native prediction without a clear gain over its original or source assembly; TabPFN's gains are uncertain.

TabPFN keeps pretrained parameters in float32 and converts and aggregates histograms in float64. Source-only checks bound the CRPS change below 10^{-4} in every application. Control risks are in Table F.9.

Context matching. Our CRUMB paper reimplementation uses 205 examples per context, 20 query clusters and 2,560 labels. Its auxiliary variants share feature–state matching geometry and

Table F.2: SFLA against original, native and source-assembly predictions. Mean CRPS gains [95% temporal-group intervals], conditional on fitted models. "Unrelated" removes state–response dependence given H; "Equal rates" removes recording shift.

TabDPT

Application	Condition	Original − SFLA	Native − SFLA	Source assembly − SFLA
Gas Turbine	Shift	-0.0003 $[-0.0044, +0.0038]$	$\mathbf{+0.0063}$ $[+0.0024, +0.0103]$	0.0000 $[-0.0016, +0.0015]$
Gas Turbine	Unrelated	$\mathbf{-0.0033}$ $[-0.0063, -0.0002]$	$+0.0030$ $[-0.0025, +0.0084]$	-0.0012 $[-0.0028, +0.0003]$
Appliances	Shift	$+0.0034$ $[-0.0003, +0.0072]$	$\mathbf{+0.0016}$ $[+0.0002, +0.0030]$	$\mathbf{+0.0007}$ $[+0.0001, +0.0014]$
Appliances	Unrelated	$+0.0022$ $[-0.0026, +0.0070]$	$+0.0013$ $[-0.0003, +0.0028]$	$+0.0006$ $[-0.0002, +0.0014]$
Beijing	Shift	$\mathbf{+0.0121}$ $[+0.0056, +0.0186]$	-0.0093 $[-0.0191, +0.0004]$	$+0.0011$ $[-0.0033, +0.0055]$
Beijing	Equal rates	$\mathbf{+0.0093}$ $[+0.0001, +0.0184]$	-0.0053 $[-0.0146, +0.0040]$	$+0.0002$ $[-0.0001, +0.0004]$
Household	Shift	$\mathbf{+0.0141}$ $[+0.0068, +0.0215]$	$\mathbf{+0.0288}$ $[+0.0197, +0.0380]$	$\mathbf{+0.0169}$ $[+0.0092, +0.0246]$
Household	Equal rates	-0.0025 $[-0.0071, +0.0021]$	$\mathbf{+0.0100}$ $[+0.0047, +0.0153]$	-0.0002 $[-0.0005, +0.0002]$
Metro	Shift	$\mathbf{+0.0228}$ $[+0.0144, +0.0312]$	$+0.0042$ $[-0.0023, +0.0107]$	$\mathbf{+0.0256}$ $[+0.0175, +0.0338]$
Metro	Equal rates	-0.0014 $[-0.0045, +0.0017]$	$\mathbf{+0.0144}$ $[+0.0069, +0.0218]$	0.0000 $[-0.0003, +0.0002]$

TabPFN 3.5

Application	Condition	Original − SFLA	Native − SFLA	Source assembly − SFLA
Gas Turbine	Shift	$+0.0005$ $[-0.0014, +0.0025]$	-0.0003 $[-0.0031, +0.0025]$	-0.0007 $[-0.0019, +0.0006]$
Gas Turbine	Unrelated	-0.0006 $[-0.0028, +0.0016]$	-0.0004 $[-0.0030, +0.0022]$	-0.0001 $[-0.0012, +0.0009]$
Appliances	Shift	-0.0005 $[-0.0029, +0.0019]$	$\mathbf{-0.0032}$ $[-0.0059, -0.0005]$	$+0.0003$ $[-0.0005, +0.0012]$
Appliances	Unrelated	$+0.0007$ $[-0.0015, +0.0030]$	-0.0010 $[-0.0026, +0.0006]$	$+0.0003$ $[-0.0002, +0.0009]$
Beijing	Shift	$\mathbf{+0.0147}$ $[+0.0089, +0.0206]$	$\mathbf{-0.0103}$ $[-0.0192, -0.0013]$	$+0.0005$ $[-0.0053, +0.0064]$
Beijing	Equal rates	$+0.0069$ $[-0.0004, +0.0141]$	$\mathbf{-0.0144}$ $[-0.0215, -0.0072]$	$+0.0001$ $[-0.0002, +0.0004]$
Household	Shift	$\mathbf{+0.0159}$ $[+0.0073, +0.0245]$	$\mathbf{+0.0130}$ $[+0.0079, +0.0181]$	$\mathbf{+0.0172}$ $[+0.0096, +0.0249]$
Household	Equal rates	-0.0015 $[-0.0084, +0.0053]$	$+0.0042$ $[-0.0024, +0.0108]$	-0.0003 $[-0.0006, +0.0001]$
Metro	Shift	$\mathbf{+0.0234}$ $[+0.0146, +0.0322]$	$\mathbf{+0.0081}$ $[+0.0030, +0.0133]$	$\mathbf{+0.0256}$ $[+0.0172, +0.0341]$
Metro	Equal rates	-0.0005 $[-0.0038, +0.0029]$	$\mathbf{+0.0241}$ $[+0.0166, +0.0316]$	-0.0001 $[-0.0004, +0.0002]$

Table F.3: CRUMB context-matching controls. Mean fair CRPS ↓ for our paper reimplementation. Auxiliary source and deployment matching use the same feature–state geometry. Household and Metro targeting improves this auxiliary construction, but its final risk remains above the original TabICLv2 predictor.

Application	Condition	CRUMB	Auxiliary: source	Auxiliary: deployment
Appliances	Shift	.1825	.1746	.1755
	Unrelated state	.1772	.1817	.1736
Household	Shift	.3931	.3965	.3863
	Equal rates	.4157	.4201	.4194
Metro	Shift	.2112	.2299	.2183
	Equal rates	.1704	.1733	.1747

change only the state probabilities (Table F.3). Household SFLA improves on ordinary CRUMB by .0256 [.0170, .0343] and auxiliary CRUMB by .0188 [.0120, .0256]. Targeting helps auxiliary CRUMB, but leaves it worse than the original by .0070 [.0012, .0127], with active-power gains and reactive-power losses. Metro SFLA improves on ordinary CRUMB by .0247 [.0159, .0334]. Both Appliances comparisons and Beijing's auxiliary update, .0069 [−.0062, .0199], remain unresolved.

Conditional means. Repeated regression and assembly improve mean prediction on Household and Metro, with a positive interval for assembly over repeated regression only on Household, driven by active power. Gas has a small harmful mixture update. Weighted Direct improves Household and harms Beijing and the Gas unrelated-state control (Tables F.4 and F.5).

This comparison implements the two-stage mean-regression algorithm of Boeken et al. (2023) with histogram gradient boosting. Both routes reuse the same state-conditional means. Assembly averages them with deployment probabilities; repeated regression predicts them on the $R = 0$ auxiliary rows and fits an outer regressor on H, without target outcomes. Weighted Direct fits a separate H-only model. RMSE evaluates these conditional means, leaving distributional extensions of repeated regression outside this comparison.

Both tree stages use squared loss, learning rate .05, minimum leaf size 20 and L_2 penalty 1. Three-fold group validation chooses between 200 iterations/15 leaves and 400 iterations/31 leaves, favoring the smaller model in ties before refitting on all permitted rows. Source fitting, Direct and outer regression use source, weighted and pseudo-response MSE, respectively. Weighted validation of source-trained trees returns the same predictions as ordinary validation in all 18 cells. The original paper primarily uses splines and also evaluates regression-tree variants.

F.1.2 Gains attributable to recording correction

Updating the mixture improves all six displayed families' source assemblies on Household and Metro. Every family also improves its Metro original. Household's gains concentrate on active power (Table F.6); for DT-FM, updating recovers only part of the loss from fitting state-specific predictors. These comparisons hold the component predictions fixed.

Household's primary Context Direct requirement failed, and SFLA was a secondary comparison. That primary test used 46 four-week bins; the common comparisons here use weekly groups. Metro's TabICLv2 and TabDPT SFLA gains exceed the prespecified .005 useful-gain threshold, but miss the separate .005 interval half-width target.

Beijing supplies a contrasting result. Weather and cyclic time predict pollution while station is withheld. Uniform resampling explains much of Context Direct's gain, and fitting the source mixture explains most of SFLA's gain. The additional targeting and mixture updates are unresolved for TabICLv2 and TabDPT. On Metro, targeting improves on uniform contexts but still loses to the original. Table F.7(a,b) reports both references. Panel (c) separates component fitting from updating using paired scores. Selection among these fixed procedures is evaluated in Appendix F.2.

Table F.4: Mean target-specific RMSE on source-fitted scales; lower is better. Bold marks all lowest displayed means in each column. RR and tree assemblies share their fitted teacher.

(a) Shift conditions

Procedure	Gas Turbine	Appliances	Beijing	Household	Metro
Trees: source fit	0.5611	0.4111	0.9443	0.7733	0.4645
Trees: weighted Direct	0.5800	**0.4090**	0.9694	0.7575	0.4538
Trees: source assembly	0.5571	0.4126	0.9417	0.7723	0.4639
Trees: repeated regression	**0.5459**	0.4268	**0.9265**	0.7558	0.4303
Trees: deployment assembly	0.5586	0.4119	0.9409	**0.7492**	**0.4234**

(b) Controls

Procedure	Gas Turbine	Appliances	Beijing	Household	Metro
Trees: source fit	0.6143	0.3988	0.8964	**0.8305**	0.3964
Trees: weighted Direct	0.6684	**0.3848**	0.9018	0.8313	0.3971
Trees: source assembly	0.6124	0.3989	0.8983	0.8350	**0.3836**
Trees: repeated regression	**0.5804**	0.4138	**0.8838**	0.8357	0.3866
Trees: deployment assembly	0.6143	0.3989	0.8984	0.8354	0.3839

Gas and Appliances use states unrelated to the response given H; Beijing, Household and Metro use equal recording rates.

Table F.5: Which operation improves mean prediction? Paired RMSE gains with pointwise 95% group delta-method t intervals, conditional on the fitted models and auxiliary sample. Positive gains favor deployment assembly in the first two columns and weighted Direct in the third. The first two comparisons share the same fitted tree teacher; Direct is a separate weighted fit.

Application	Condition	Mixture updating gain	Assembly gain over RR	Weighted Direct gain over source fit
		-0.00145	-0.01269	-0.01886
Gas Turbine	Shift	$[-0.00261, -0.00029]$	$[-0.04560, +0.02023]$	$[-0.05235, +0.01462]$
		-0.00193	-0.03391	-0.05410
Gas Turbine	Unrelated state	$[-0.00473, +0.00088]$	$[-0.06949, +0.00167]$	$[-0.09459, -0.01362]$
		$+0.00070$	$+0.01489$	$+0.00208$
Appliances	Shift	$[+0.00008, +0.00131]$	$[-0.01659, +0.04637]$	$[-0.01121, +0.01536]$
		$+0.00006$	$+0.01493$	$+0.01397$
Appliances	Unrelated state	$[-0.00006, +0.00018]$	$[-0.01660, +0.04647]$	$[+0.00303, +0.02491]$
		$+0.00081$	-0.01438	-0.02510
Beijing	Shift	$[-0.00343, +0.00504]$	$[-0.03542, +0.00666]$	$[-0.04553, -0.00467]$
		-0.00011	-0.01458	-0.00539
Beijing	Equal rates	$[-0.00037, +0.00014]$	$[-0.03933, +0.01018]$	$[-0.01340, +0.00261]$
		$+0.02313$	$+0.00663$	$+0.01583$
Household	Shift	$[+0.01095, +0.03530]$	$[+0.00073, +0.01254]$	$[+0.00232, +0.02934]$
		-0.00039	$+0.00032$	-0.00087
Household	Equal rates	$[-0.00100, +0.00022]$	$[-0.00700, +0.00765]$	$[-0.00389, +0.00214]$
		$+0.04047$	$+0.00695$	$+0.01065$
Metro	Shift	$[+0.02457, +0.05638]$	$[-0.00290, +0.01679]$	$[-0.00061, +0.02192]$
		-0.00030	$+0.00272$	-0.00077
Metro	Equal rates	$[-0.00100, +0.00040]$	$[-0.00625, +0.01169]$	$[-0.00282, +0.00127]$

Table F.6: Household recording shift, separated by response. Mean CRPS ↓; bold marks the smallest displayed value in each row. Active power drives the mean mixture gain across the three pretrained models.

Model	Response	Original	Source assembly	SFLA
TabICLv2	Active power	0.3111	0.3188	**0.2873**
	Reactive power	**0.4475**	0.4477	0.4477
TabDPT	Active power	0.3113	0.3166	**0.2832**
	Reactive power	0.4503	0.4506	**0.4501**
TabPFN 3.5	Active power	0.3130	0.3167	**0.2815**
	Reactive power	0.4498	**0.4488**	0.4495

Table F.7: Frozen-predictor comparisons: mean CRPS gains [95% temporal-group intervals], conditional on fits. (a,b) Each row names the reference and adapted procedure. (c) TabICLv2's gain from fitting state-specific predictions and updating their probabilities, computed from paired scores. Gas/Appliances controls use unrelated states; the others equalize rates.

(a) Recording shift

CRPS gain ↑	Gas Turbine	Appliances	Beijing	Household	Metro
TabICLv2					
Original − Context Direct	−.0036 [−.0241, .0169]	−.0043 [−.0122, .0036]	.0334 [.0183, .0484]	−.0025 [−.0079, .0029]	−.0071 [−.0138, −.0004]
Original − SFLA	.0024 [−.0027, .0075]	.0027 [−.0006, .0060]	.0068 [.0024, .0111]	.0118 [.0045, .0191]	.0233 [.0145, .0320]
Uniform − Context Direct	−.0083 [−.0225, .0058]	−.0090 [−.0206, .0025]	.0092 [−.0014, .0197]	.0068 [.0019, .0117]	.0062 [.0003, .0121]
TabDPT					
Original − Context Direct	−.0022 [−.0130, .0087]	−.0061 [−.0145, .0022]	.0436 [.0276, .0595]	−.0046 [−.0084, −.0008]	−.0075 [−.0105, −.0044]
Original − SFLA	−.0003 [−.0044, .0038]	.0034 [−.0003, .0072]	.0121 [.0056, .0186]	.0141 [.0068, .0215]	.0228 [.0144, .0312]
Uniform − Context Direct	−.0048 [−.0126, .0031]	−.0087 [−.0205, .0031]	.0049 [−.0049, .0147]	−.0034 [−.0075, .0007]	.0024 [−.0002, .0050]

(b) Controls

CRPS gain ↑	Gas Turbine	Appliances	Beijing	Household	Metro
TabICLv2					
Original − Context Direct	−.0033 [−.0135, .0070]	−.0007 [−.0096, .0082]	+0.0330 [+0.0191, +0.0470]	−.0095 [−.0126, −.0064]	−.0225 [−.0282, −.0169]
Original − SFLA	.0007 [−.0021, .0036]	.0022 [−.0005, .0049]	+0.0057 [+0.0017, +0.0097]	−.0018 [−.0067, .0031]	−.0014 [−.0044, .0015]
Uniform − Context Direct	−.0011 [−.0107, .0085]	−.0030 [−.0097, .0037]	+0.0050 [−0.0033, +0.0133]	.0013 [−.0009, .0034]	−.0035 [−.0057, −.0013]
TabDPT					
Original − Context Direct	−.0027 [−.0099, .0044]	−.0003 [−.0069, .0062]	+0.0358 [+0.0189, +0.0527]	−.0043 [−.0072, −.0015]	−.0062 [−.0092, −.0033]
Original − SFLA	−.0033 [−.0063, −.0002]	.0022 [−.0026, .0070]	+0.0093 [+0.0001, +0.0184]	−.0025 [−.0071, .0021]	−.0014 [−.0045, .0017]
Uniform − Context Direct	−.0077 [−.0143, −.0010]	−.0045 [−.0098, .0007]	+0.0034 [−0.0040, +0.0107]	.0013 [−.0003, .0029]	−.0013 [−.0025, −.0000]

(c) Fitting and updating the TabICLv2 mixture

Application	Original − source assembly	Source assembly − SFLA
Gas Turbine	+0.0027 [-0.0027, +0.0080]	-0.0003 [-0.0013, +0.0008]
Appliances	+0.0020 [-0.0013, +0.0053]	+0.0007 [+0.0002, +0.0012]
Beijing	+0.0065 [+0.0016, +0.0113]	+0.0003 [-0.0033, +0.0038]
Household	-0.0040 [-0.0082, +0.0002]	+0.0158 [+0.0082, +0.0234]

Table F.8: Exact histogram CRPS for the fixed TabDPT predictions on Gas and Appliances. The same fitted distributions and cases underlie the 128-draw main risks. Gains average the two responses before paired group inference.

Application	Original	Source assembly	SFLA	Update [95% interval]
Gas Turbine	0.19016	0.19136	0.19173	-0.00036 $[-0.00220, +0.00147]$
Appliances	0.18303	0.18121	0.18067	$+0.00054$ $[-0.00002, +0.00109]$

F.1.3 Recording designs and limits

Recording must change the state composition, and the states must predict different response distributions after conditioning on H. For a binary state, the population opportunity is

$$\Delta^{\ell}_{\text{epi}} = \mathbb{E}_{H\sim P_0}\left[\{\pi_0(1 \mid H) - \pi_1(1 \mid H)\}^2 D_\ell\{Q_F(\cdot \mid H, 0), Q_F(\cdot \mid H, 1)\}\right]. \tag{F.1}$$

For CRPS, $D_\ell = \int\{F_1(y \mid h) - F_0(y \mid h)\}^2 dy$, with F_c the true state-conditional response CDF. The unrelated-state control removes this separation; equal recording rates remove the probability change. Both give zero population opportunity. On measured data the true components are unknown, and a fitted risk difference also contains estimation error.

Recording rates determine the direction of the composition change. With $e_c = \Pr(R = 1 \mid H = h, C = c)$ and $p_r = \Pr(C = 1 \mid H = h, R = r)$, Bayes' rule gives

$$\frac{p_0/(1-p_0)}{p_1/(1-p_1)} = \frac{e_0(1-e_1)}{e_1(1-e_0)}. \tag{F.2}$$

Thus deployment increases Saturday's odds by $49/9$ on Metro and reduces active-submeter odds by $9/49$ on Household. The probability change can still be small when H already predicts the state. Gas and Appliances assign state through a logistic function of the product of residualized responses. Under independent symmetric residuals this can change dependence without changing marginal distributions. The measured residuals need not satisfy that condition, and their marginal CRPS results alone cannot settle joint accuracy.

Controls without a conditional change. Beijing's equal-rate control assigns both stations recording probability $.500205$, preserving the expected recording rate, 2,560-label budget, 4,096 auxiliary records and 2,048 evaluation cases. SFLA still improves the TabICLv2 and TabDPT originals, almost entirely through source assembly. Four-week grouping retains TabICLv2's positive total gain but lowers TabDPT's interval bound to $-.00003$. Shifted and equal-rate allocations contain different cases and are evaluated separately. Equal-rate Household and Metro controls have near-zero mixture updates, even when replacing native missing-feature handling improves accuracy. All five TabPFN control intervals against its original include zero. TabDPT retains its Gas unrelated-state loss. TabPFN context targeting also harms that control relative to uniform contexts, by $.00535$ $[.00003, .01066]$.

The exact histogram updates in Table F.8 are used for both marked main-table contrasts. On Appliances, the historical 128-draw gain is $.00073$ $[.00009, .00138]$; exact integration gives $.00054$ $[-.00002, .00109]$, leaving improvement unresolved. The Gas control loss persists. Later head studies score a 512-atom approximation exactly, which is a different predictive representation.

Coverage remains imperfect: TabPFN SFLA reaches 85.4% on Metro and 81.1% on Appliances at nominal 90%. These comparisons evaluate risk and coverage separately.

Table F.9: Native prediction and recording correction under the two controls: mean CRPS ↓. All methods predict without the individual's C. Gas and Appliances use an unrelated state; the others equalize recording rates. Bold numbers mark the lowest displayed means within each row, including ties. Shifted risks: Table 3.

Model	Application	Original	Native	Uniform	Context Direct	Source assembly	SFLA
TabDPT	Gas Turbine	0.2086	0.2148	**0.2036**	0.2113	0.2106	0.2118
	Appliances	0.1826	0.1817	**0.1784**	0.1830	0.1810	0.1805
	Beijing	0.4495	0.4349	0.4170	**0.4136**	0.4403	0.4402
	Household	**0.4043**	0.4168	0.4099	0.4086	0.4066	0.4068
	Metro	**0.1600**	0.1757	0.1649	0.1662	0.1613	0.1614
TabPFN 3.5	Gas Turbine	0.1989	0.1990	**0.1978**	0.2032	0.1993	0.1994
	Appliances	0.1699	**0.1681**	0.1763	0.1786	0.1695	0.1691
	Beijing	0.4409	**0.4196**	0.4325	0.4254	0.4341	0.4340
	Household	0.4070	0.4128	0.4047	**0.4042**	0.4083	0.4086
	Metro	0.1629	0.1874	**0.1626**	0.1628	0.1632	0.1633

F.2 Learning and selecting pretrained predictions

Recording weights can improve a predictor through its learning objective or through the choice of which fitted prediction to return. We test both routes and compare their gains with using the same labels as additional demonstrations. Fitting and selection labels are separate from evaluation. CRPS is exact for the finite predicted distributions, whereas the main application table estimates CRPS from 128 draws. Unless specified otherwise, intervals are paired, pointwise 95% temporal-group intervals conditional on fitted models. Electricity uses an independent allocation. The other studies use records examined during development.

F.2.1 The effect of weighting on learning

On Household active power, weighting both training and stopping improves TabICLv2, while changing either alone retains initialization (Table F.10a). The gain is 0.0271 $[0.0148, 0.0393]$, with interaction 0.0123 $[0.0025, 0.0222]$. The two 30-epoch trajectories share initialization, labels, contexts and episodes, and differ in source versus weighted pinball loss. Source MSE, source CRPS and weighted CRPS select from each trajectory, including epoch zero. Validation labels remain outside inference contexts, and all scores use the same eight-member predictive distribution.

The benefit varies across outcomes and recording conditions. Extending training to 100 epochs selects epoch 20 with a similar Household gain. The two Beijing gains remain unresolved, and Household without shift loses 0.0104 $[0.0022, 0.0186]$ to the full-label original (Table F.10b). Spending labels on learning and validation can therefore reduce accuracy relative to using them as demonstrations.

Comparison with native fine-tuning. The released pinball recipe uses learning rate 10^{-5}, at most 30 epochs and source-MSE stopping with patience eight and epoch zero eligible. Training and validation split 2,048 labels, with 512 separate scaling labels. Native inference uses training contexts. A full-context control adds validation labels to both frozen and tuned contexts after stopping. Four models update their parameters, while Household reactive power retains initialization. Beijing gains remain uncertain, and Metro and Household active power worsen under both context rules (Table F.10c). Preprocessing, ensemble size and output format agree within each comparison.

On Gas, weighted training improves initialization by 0.0034 $[0.0010, 0.0058]$ but adds 0.0000 $[-0.0042, 0.0042]$ over source training. The corresponding unrelated-state gains are 0.0066 $[0.0021, 0.0111]$ and 0.0011 $[-0.0018, 0.0040]$. Both learners share initialization, group-separated episodes, retained training contexts, 2,048 labels and 512 scaling labels. Both use weighted validation CRPS with initialization eligible. Intervals average both responses over 141/143 groups. Training helps without a resolved weighting benefit.

Weighting an output head. On Household, weighted head fitting improves average risk for both backbones without changing contexts or pretrained outputs. Active-power gains outweigh reactive-

Table F.10: Training and stopping for TabICLv2. (a) CRPS ↓ and selected epochs, from initial risk 0.3120. (b) Eight checkpoints over 100 epochs, with a full-label original using all 2,048 fitting/stopping labels as contexts. Both studies reserve 512 scaling labels. (c) Frozen-minus-tuned CRPS for five native 30-epoch fits evaluated with training-only and full contexts. Positive gains favor tuning. Beijing and Household intervals use 60 and 80 weeks.

(a) Household active power: training and stopping

Stopping loss	Source training	Weighted training	Weighting gain [95% interval]
Source mean squared error	0.3267 (30)	0.3120 (0)	$+\mathbf{0.0147}$ [+0.0090, +0.0204]
Source CRPS	0.3267 (27)	0.3120 (0)	$+\mathbf{0.0147}$ [+0.0090, +0.0204]
Weighted CRPS	0.3120 (0)	0.2849 (14)	$+\mathbf{0.0271}$ [+0.0148, +0.0393]

(b) Transfer and no-shift control

Target	Initial	Source training source CRPS	Native patience	Direct	Full-label original
Beijing $PM_{2.5}$	0.4704	0.4884	0.4703	0.4693	0.4635
Beijing PM_{10}	0.4824	0.4832	0.4806	0.4804	0.4684
Household, no shift	0.3446	0.3456	0.3449	0.3462	0.3357

(c) Native fine-tuning and inference contexts

Target	Epoch	Training-only contexts	Full contexts after stopping
Beijing $PM_{2.5}$	2	+0.00008 [−0.00075, +0.00091]	−0.00003 [−0.00088, +0.00082]
Beijing PM_{10}	6	+0.00279 [−0.00418, +0.00975]	−0.0005 [−0.00764, +0.00663]
Metro traffic	2	$-\mathbf{0.00071}$ [−0.0009, −0.00052]	$-\mathbf{0.00069}$ [−0.00087, −0.00051]
Household active	30	$-\mathbf{0.0151}$ [−0.0209, −0.00928]	$-\mathbf{0.0138}$ [−0.0180, −0.00956]
Household reactive	0	+0.0000 [+0.0000, +0.0000]	+0.0000 [+0.0000, +0.0000]

power losses (Table F.11). The comparison uses the same four-region CDF head for source and weighted CRPS fitting (Appendix C.2). Five weeks supply 122 or 128 head labels beyond 2,048 context and 512 scaling labels. Evaluation scores 512-midpoint-quantile distributions over 80 weeks, grouped into 46 four-week bins. Without shift, original-to-weighted-head risks are .4079 → .4078 for TabICLv2 and .4050 → .4056 for TabDPT.

Table F.11: Household head learning with fixed original contexts. (a) Reference-minus-weighted-head CRPS, averaged over responses. (b) Gains over the original for each response. Brackets give 95% four-week-group intervals conditional on predictions and fitted heads.

(a) Mean gains over the stated reference

Reference	TabICLv2	TabDPT
Original	**0.0115** [0.0012,0.0218]	**0.0128** [0.0011,0.0245]
Source-loss head	**0.0157** [0.0036,0.0277]	**0.0205** [0.0070,0.0340]
Equal-label original	**0.0121** [0.0011,0.0232]	**0.0138** [0.0015,0.0262]
SFLA	-0.0003 [-0.0054,0.0048]	-0.0013 [-0.0077,0.0051]

(b) Response-specific gains over the original

Response	Predictor	Original	Weighted head	Gain [95% interval]
Active power	TabICLv2	0.3108	0.2837	**0.0271** [0.0084,0.0458]
Active power	TabDPT	0.3117	0.2799	**0.0318** [0.0110,0.0526]
Reactive power	TabICLv2	0.4480	0.4522	**-0.0042** [-0.0080,-0.0004]
Reactive power	TabDPT	0.4503	0.4566	**-0.0063** [-0.0114,-0.0012]

Across applications, ordinary head fitting explains most of Beijing's gain, weighting helps Metro/TabDPT relative to its source head without a clear gain over the original, and both comparisons worsen on Appliances (Figure F.1). These results use the selection allocations described below. Their Household allocation differs from Table F.11 and has unresolved weighting increments. The independent Electricity test also gives unresolved primary-shift weighting gains: −0.00022

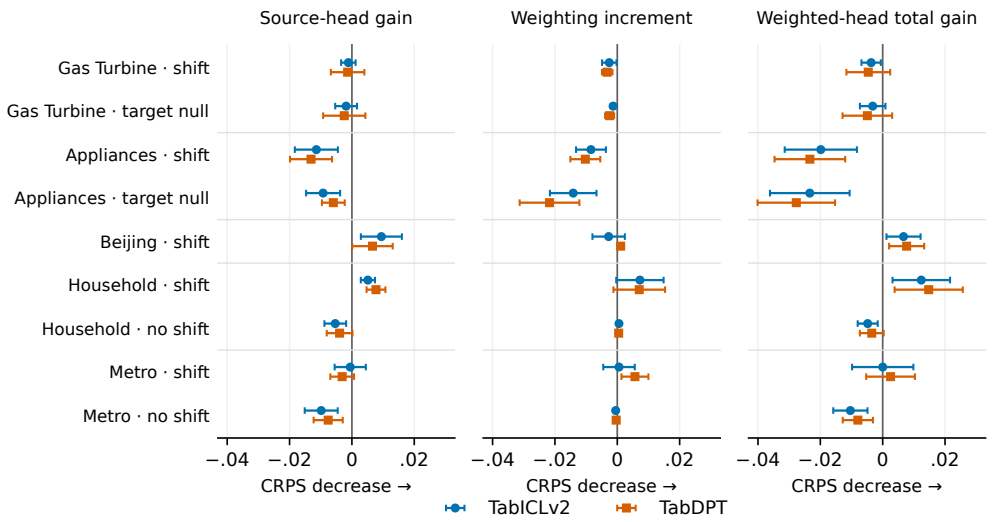


Figure F.1: Head learning across applications. Source-head gain plus the weighting increment equals total weighted-head gain. Points and bars show mean CRPS gains and 95% temporal-group intervals, conditional on fits and averaged over responses. "Target null" denotes unrelated recording states; "no shift" denotes equal recording rates.

$[-0.00047, 0.00019]$ and 0.00009 $[-0.00013, 0.00039]$. Reversing the rates helps TabICLv2 relative to its source head, while both heads remain worse than the original.

F.2.2 Recording correction in the selected predictor

The fixed library contains the original predictor, the average of two uniform contexts, SFLA and a weighted head on original contexts. Weighted or ordinary source CRPS selects one member, breaking ties in that order. The budget is 2,048 context labels, 512 scaling labels and 256 further labels split by group between head fitting and selection. Heads remain fixed after selection. Validation contains five weeks on Household and 103–141 labels in 5–22 groups elsewhere. Risks average responses within cases. Household intervals use 46 four-week evaluation groups; its equal-label original receives all 2,816 labels.

Selected predictions improve on originals on Beijing, Metro and shifted Household, with mixed results on Gas and Appliances (Table F.12). Beijing returns uniform contexts under both criteria. Household returns heads for active power and SFLA for reactive power, with no clear gain from weighting selection. Without shift, TabDPT chooses uniform contexts and loses to its equal-label original. Weighting selection also harms Appliances/TabICLv2. Selection can also lose to fixed SFLA: TabICLv2's gains over it are -0.0048 $[-0.0095, -0.0002]$ on Gas with unrelated states and -0.0069 $[-0.0124, -0.0014]$ on shifted Appliances. Thus some losses reflect which candidate validation returns, even with a useful SFLA prediction available.

Contribution of mixture updating. Replacing source assembly by SFLA in the library holds components, heads, labels, scoring and tie-breaking fixed. On Metro, weighted selection returns SFLA for both backbones and improves on the source-library choice (Table F.14a). It also improves equal-label originals by 0.02253 $[0.01414, 0.03091]$ and 0.02220 $[0.01392, 0.03048]$. Beijing keeps uniform contexts, and Household keeps the same weighted head for active power: correcting a candidate can leave the selected prediction unchanged. Substitution gives zero or unresolved gains on Gas, Appliances and the equal-rate controls. Metro's gains are unresolved under ordinary selection.

Weighting validation with native prediction available. Weighted validation is compared with ordinary validation on five fixed TabPFN 3.5 candidates: Original, Native, Uniform, Context Direct and SFLA. Both criteria score the same predictions and break ties in that order. All use the same frozen checkpoint, native float32 inference and float64 histogram aggregation. Uniform and Context Direct

Table F.12: Weighted selection: CRPS gains over the original and over selection by unweighted source CRPS. Positive values favor weighted selection. Brackets give paired, pointwise 95% temporal-group intervals conditional on fits and choices. Exact zeros denote identical predictions.

Application / condition	Model	Gain over original	Gain over ordinary selection
Household: original receives the full adaptation label budget			
Household: shift	TabICLv2	$+0.0119$ $[+0.0023, +0.0215]$	$+0.0018$ $[-0.0026, +0.0061]$
Household: shift	TabDPT	$+0.0147$ $[+0.0045, +0.0250]$	$+0.0019$ $[-0.0030, +0.0068]$
Household: equal rates	TabICLv2	-0.0004 $[-0.0020, +0.0012]$	0 $[0, 0]$
Household: equal rates	TabDPT	-0.0054 $[-0.0079, -0.0029]$	0 $[0, 0]$
Other applications: original uses 2,048 context and 512 scaling labels			
Beijing	TabICLv2	$+0.0256$ $[+0.0121, +0.0390]$	0 $[0, 0]$
Beijing	TabDPT	$+0.0384$ $[+0.0237, +0.0530]$	0 $[0, 0]$
Metro	TabICLv2	$+0.0234$ $[+0.0148, +0.0321]$	$+0.0234$ $[+0.0148, +0.0321]$
Metro	TabDPT	$+0.0235$ $[+0.0151, +0.0319]$	$+0.0235$ $[+0.0151, +0.0319]$
Metro: no shift	TabICLv2	0 $[0, 0]$	0 $[0, 0]$
Metro: no shift	TabDPT	0 $[0, 0]$	0 $[0, 0]$
Gas Turbine	TabICLv2	$+0.0005$ $[-0.0029, +0.0038]$	0 $[0, 0]$
Gas Turbine	TabDPT	-0.0007 $[-0.0035, +0.0021]$	0 $[0, 0]$
Gas: unrelated state	TabICLv2	-0.0043 $[-0.0094, +0.0008]$	0 $[0, 0]$
Gas: unrelated state	TabDPT	-0.0013 $[-0.0036, +0.0009]$	$+0.0011$ $[-0.0011, +0.0032]$
Appliances	TabICLv2	-0.0038 $[-0.0089, +0.0013]$	-0.0065 $[-0.0120, -0.0009]$
Appliances	TabDPT	-0.0005 $[-0.0041, +0.0030]$	$+0.0050$ $[-0.0003, +0.0102]$
Appliances: unrelated state	TabICLv2	$+0.0019$ $[-0.0025, +0.0063]$	0 $[0, 0]$
Appliances: unrelated state	TabDPT	-0.0006 $[-0.0051, +0.0038]$	0 $[0, 0]$

average two context bags. Source assembly is evaluated only for attribution. The total budget is 2,560 source labels, including 512 for scaling. Full-context Original and Native use the remaining 2,048 as demonstrations. The candidates reserve one eighth of source groups for validation: 7/12/9 groups on Household/Metro/Beijing. They retain these same contexts at deployment. A one-quarter allocation tests sensitivity to validation support. Choices are fixed before scoring deployment outcomes on these previously examined allocations.

Weighting replaces Uniform with SFLA for Household active power and with Context Direct on Metro, improving both application averages (Table F.13). The selected predictor also improves on Household's full-context Native, driven by active power. Reactive power has gain -0.0023 $[-0.0057, 0.0011]$. Under shift, Metro and Beijing give no resolved gain over Native. Including Native improves Beijing under shift by 0.0063 $[0.0011, 0.0114]$, also when it replaces Original to keep the library size fixed. Improving on an H-only original therefore leaves an important comparison with native prediction.

Two limitations remain visible at the response level. Metro selects Context Direct even though its fitted SFLA candidate has lower evaluation CRPS by 0.0062 $[0.0004, 0.0120]$. This comparison diagnoses selection error without replacing the source-selected prediction. Under equal recording rates, both selectors choose SFLA for Beijing PM2.5 and lose 0.0177 $[0.0030, 0.0323]$ to full-context Native. Label allocation is another cost: withholding validation groups increases Household's no-shift Original risk by 0.0043 $[0.0015, 0.0070]$. The larger validation allocation improves some comparisons and worsens others.

For the application averages, target scores are averaged within cases before calculating intervals over 80/99/60 temporal groups on Household/Metro/Beijing. Grouping adjacent evaluation periods more coarsely retains the positive Household and Metro weighting gains and Household's Native-relative gain. The intervals measure evaluation uncertainty conditional on the fitted predictions and source choices. They exclude variation from repeating context construction and selection on new source samples.

Table F.13: TabPFN 3.5 selection at a fixed total label budget. Full-context references use labels reserved for validation by the candidate predictors. Scores are exact histogram CRPS; bold marks the lowest displayed risk in each row. Positive gains favor weighted selection; zeros indicate identical predictions. Intervals are paired, pointwise 95% group intervals conditional on fitted contexts and correction estimates.

(a) Predictive risk: CRPS ↓

	Full-context references		Validation criterion	
Application / condition	Original	Native	Unweighted	Weighted
Household: shift	0.3804	0.3771	0.3819	**0.3685**
Household: equal rates	**0.4069**	0.4123	0.4112	0.4112
Metro: shift	0.2110	0.1957	0.2094	**0.1939**
Metro: equal rates	**0.1632**	0.1868	0.1642	0.1642
Beijing: shift	0.4668	**0.4432**	0.4487	0.4487
Beijing: equal rates	0.4419	**0.4196**	0.4295	0.4295

(b) Gains of weighted selection: mean [95% interval]

Application / condition	Over full-context Original	Over full-context Native	Over unweighted selection
Household: shift	$+0.0119$ $[+0.0023, +0.0214]$	$+0.0085$ $[+0.0024, +0.0146]$	$+0.0134$ $[+0.0066, +0.0202]$
Household: equal rates	-0.0043 $[-0.0107, +0.0022]$	$+0.0011$ $[-0.0058, +0.0081]$	0 $[0, 0]$
Metro: shift	$+0.0171$ $[+0.0124, +0.0217]$	$+0.0018$ $[-0.0026, +0.0061]$	$+0.0155$ $[+0.0114, +0.0196]$
Metro: equal rates	-0.0010 $[-0.0022, +0.0002]$	$+0.0226$ $[+0.0152, +0.0300]$	0 $[0, 0]$
Beijing: shift	$+0.0182$ $[+0.0032, +0.0332]$	-0.0055 $[-0.0190, +0.0080]$	0 $[0, 0]$
Beijing: equal rates	$+0.0124$ $[-0.0014, +0.0261]$	-0.0099 $[-0.0244, +0.0046]$	0 $[0, 0]$

Independent evaluation. Electricity tests the four-candidate TabICLv2/TabDPT procedure on independent data, varying validation support from five to 64 source clients. It predicts log hourly consumption from one-hour/day/week lags and cyclic hour/month features, withholding weekend status. Recording rates are $0.8/0.2$, reversed to $0.2/0.8$, or equal at $0.5/0.5$. Client groups separate fitting, validation, auxiliary and evaluation data. Each model uses 2,048 context and 256 head labels; the larger validation set contains 1,280 labels. Equal-label originals receive head and validation labels as demonstrations.

Both backbones return SFLA with 64 validation clients under the primary shift. All four primary gains over original and uniform-context predictors fail the 0.01 useful-gain requirement under Bonferroni-adjusted one-sided tests, and even the two-sided upper limits lie below 0.01. Gains over equal-label originals remain unresolved (Table F.14b). More validation data changes some choices without meeting the requirement. Weighted and ordinary selection agree in all six cases, and TabICLv2 loses to its equal-label original without shift. This independent test limits the claim that a fixed procedure reliably improves prediction.

In an exploratory expansion with Context Direct and additional heads, the best evaluation-scored candidate improves only on Beijing, by about 0.004. Metro/TabICLv2 adds no better candidate but incurs 0.0316 $[0.0230, 0.0402]$ additional selected risk. Enlarging the library can therefore increase selection error.

Table F.14: Correction after selection and independent evaluation. (a) Source-assembly-minus-SFLA CRPS before and after weighted selection on previously examined allocations, using 2,816 labels. (b) Independent Electricity: gains of the 64-client choice. Exact CRPS scores 512-atom predictions on 114 clients across 43 weeks. Its intervals resample clients and two-week blocks; one/four-week blocks provide sensitivity checks. Both panels report paired 95% intervals conditional on the fitted predictions. The two panels use distinct data and comparisons.

(a) Mixture correction before and after selection

Application	Backbone	Fixed gain	Selected gain [95% interval]
Beijing	TabICLv2	0.00099	0 [0, 0]
	TabDPT	0.00247	0 [0, 0]
Household	TabICLv2	0.01584	0.00000 [−0.00010, 0.00009]
	TabDPT	0.01699	0.00013 [−0.00012, 0.00038]
Metro	TabICLv2	0.02558	**0.02558** [0.01716, 0.03400]
	TabDPT	0.02542	**0.02352** [0.01508, 0.03195]

(b) Independent Electricity test

Condition	Model	Gain over original	Gain over equal-label original	Gain over 5-group choice
Recording shift	TabICLv2	**+0.0003** [+0.00008, +0.0006]	−0.0012 [−0.0028, +0.0015]	**+0.0033** [+0.0012, +0.0052]
Recording shift	TabDPT	**+0.0005** [+0.0001, +0.0007]	−0.0010 [−0.0025, +0.0011]	0 [0, 0]
Reverse shift	TabICLv2	+0.0003 [−0.0003, +0.0006]	−0.0013 [−0.0026, +0.0003]	**+0.0013** [+0.0005, +0.0020]
Reverse shift	TabDPT	0 [0, 0]	−0.0007 [−0.0015, +0.0002]	**+0.0009** [+0.0001, +0.0017]
No shift	TabICLv2	+0.00007 [−0.0002, +0.0004]	**−0.0037** [−0.0062, −0.0003]	+0.0005 [−0.0006, +0.0018]
No shift	TabDPT	−0.00001 [−0.0001, +0.00008]	−0.0022 [−0.0039, +0.00002]	0 [0, 0]

F.3 Flow validation and recording correction

Both flow families use one common configuration across the five shifted applications and four controls: unrelated states on Gas/Appliances and equal rates on Household/Metro. Each family fits an H-only original, weighted Direct and a state-conditional model reused by source assembly and SFLA. Fits use 2,048 training, 512 scaling and 256 validation labels, 32,768 updates and 512 midpoint steps. Within a family, initialization and input dimensions agree. The H-only model holds the state input constant, and Direct normalizes training weights by their mean. Architectures and optimization differ between families (Table C.3).

Weighted source energy selects among updates 256, 1,024, 8,192 and 32,768 using the earliest-within-one-standard-error rule. Selected and terminal predictions share the training trajectory, sampler, cases and random numbers. DT-FM chooses the first three checkpoints in 15, ten and two fits. CFMI chooses 256 in 13 fits and 1,024 in 14. Every selected checkpoint precedes the final one, although the full training trajectory counts toward computation. Resolution checks compare 512 and 1,024 steps.

Correction after validation. Selecting a checkpoint improves both families. All four DT-FM prediction rules improve on Gas, Beijing, Household and Metro in both CRPS and joint energy; Appliances changes are uncertain. Table F.15 therefore measures correction against selected originals. Both flow assemblies lose on Beijing after fitting state-specific predictors. CFMI gains on Household through both fitting and updating. DT-FM's Metro gain comes mainly from updating, while its Household net change is uncertain.

Joint updating gains. Joint energy gives the same direction of mixture-update findings as marginal CRPS on Gas and Appliances. On the same selected predictions, Gas updating gains are -0.0046 $[-0.0084, -0.0009]$ for DT-FM and -0.0003 $[-0.0038, 0.0031]$ for CFMI. On Appliances they are $+0.0059$ $[0.0006, 0.0112]$ and $+0.0029$ $[-0.0015, 0.0073]$. These paired temporal-group intervals use energy divided by the square root of target dimension. DT-FM's Gas loss therefore persists under joint scoring. All four corresponding independent-state intervals include zero.

Scoring and calibration. Selection minimizes joint energy divided by $\sqrt{|T|}$. Reported CRPS averages the same targets. Evaluation uses 128 draws per component, with 32/64-draw checks and the paired-component correction in Appendix C.5. Temporal-group intervals condition on fitted models. Selection improves coverage, but marked undercoverage remains on Appliances (Table F.17).

Table F.15: Recording correction after flow validation. Mean CRPS gains [95% temporal-group intervals], conditional on fits. Columns separate fitting state-specific predictors, updating their probabilities and the total change from the selected original.

Application	Selected flow	Original − source assembly	Source assembly − SFLA	Original − SFLA
Gas Turbine	DT-FM	$+0.0066$ $[-0.0074, +0.0206]$	$\mathbf{-0.0039}$ $[-0.0071, -0.0008]$	$+0.0026$ $[-0.0119, +0.0172]$
Gas Turbine	CFMI	$\mathbf{+0.0113}$ $[+0.0014, +0.0212]$	-0.0005 $[-0.0034, +0.0023]$	$\mathbf{+0.0108}$ $[+0.0006, +0.0210]$
Appliances	DT-FM	$+0.0025$ $[-0.0115, +0.0165]$	$\mathbf{+0.0054}$ $[+0.0003, +0.0105]$	$+0.0079$ $[-0.0040, +0.0198]$
Appliances	CFMI	-0.0033 $[-0.0120, +0.0053]$	$+0.0031$ $[-0.0012, +0.0074]$	-0.0002 $[-0.0081, +0.0077]$
Beijing	DT-FM	$\mathbf{-0.0202}$ $[-0.0359, -0.0046]$	$+0.0038$ $[-0.0011, +0.0087]$	$\mathbf{-0.0164}$ $[-0.0315, -0.0012]$
Beijing	CFMI	$\mathbf{-0.0211}$ $[-0.0342, -0.0080]$	$+0.00005$ $[-0.0030, +0.0031]$	$\mathbf{-0.0210}$ $[-0.0337, -0.0083]$
Household	DT-FM	$\mathbf{-0.0086}$ $[-0.0137, -0.0035]$	$+0.0055$ $[-0.0010, +0.0120]$	-0.0031 $[-0.0121, +0.0059]$
Household	CFMI	$\mathbf{+0.0264}$ $[+0.0147, +0.0380]$	$\mathbf{+0.0228}$ $[+0.0144, +0.0312]$	$\mathbf{+0.0492}$ $[+0.0344, +0.0640]$
Metro	DT-FM	$+0.0003$ $[-0.0075, +0.0080]$	$\mathbf{+0.0275}$ $[+0.0190, +0.0359]$	$\mathbf{+0.0277}$ $[+0.0210, +0.0345]$
Metro	CFMI	$+0.0032$ $[-0.0011, +0.0074]$	$\mathbf{+0.0118}$ $[+0.0046, +0.0190]$	$\mathbf{+0.0150}$ $[+0.0071, +0.0229]$

Table F.16: Selected-flow CRPS under unrelated-state and equal-rate controls. Gas and Appliances retain recording imbalance but remove state–response dependence; Household and Metro equalize recording rates. These flow comparisons do not include a Beijing control.

Application	Selected flow	Original	Weighted Direct	Source assembly	SFLA
Gas Turbine	DT-FM	0.2489	0.2775	<u>0.2474</u>	0.2501
Gas Turbine	CFMI	0.2585	0.2584	<u>0.2499</u>	0.2513
Appliances	DT-FM	<u>0.2231</u>	0.2416	0.2287	0.2289
Appliances	CFMI	0.2380	<u>0.2076</u>	0.2332	0.2347
Household	DT-FM	0.4456	0.4461	<u>0.4419</u>	0.4422
Household	CFMI	0.4560	0.4551	<u>0.4422</u>	0.4423
Metro	DT-FM	0.1813	0.1814	0.1723	<u>0.1721</u>
Metro	CFMI	0.1852	0.1904	<u>0.1788</u>	0.1788

Validation criterion and fitting variation. All validation rules use the same 54 trajectories, source rows and checkpoint grid. Weighted or ordinary endpoint energy and held-out flow-matching loss are compared under minimum-loss and one-standard-error selection. Both losses select predictions substantially better than the terminal checkpoint. Their differences are usually unresolved, with the remaining intervals favoring different criteria. The close means in Table F.17b therefore support useful validation with either loss, without establishing equivalence. All 18 individual selection gains have positive temporal-group intervals. Selection also improves coverage, while substantial undercoverage remains on Appliances.

Appendix D.3.1 compares fitted joint predictions with the attainable oracle under an unchanged-marginals construction.

F.4 Conditional completion

Across all eight queries, selected DT-FM and CFMI have lower mean energy than both DiffPuter rules. Adaptive MIWAE leads on California's larger query. Original and selected samplers are shown together to expose the contribution of checkpoint choice.

California, Protein, SGEMM and Tetouan each supply three source fits. Source pools contain 4,096 rows, or 4,176 on Tetouan. Evaluation uses 512 rows and 64 predictive draws. Each fit serves two specified completion blocks: two/four coordinates for California, Protein and Tetouan, and one/two for SGEMM. The remaining values are observed. All source records are complete, so recording correction is unnecessary. Artificial masks define the training and evaluation tasks.

DT-FM and CFMI use unweighted flow matching with 32,768 and 75,000 updates, respectively (Table C.3). Source validation selects checkpoints. CFMI also increases Euler steps from 100 to 400. A matched Tetouan crossing isolates these changes: at 400 steps, checkpoint selection gains 0.1094 energy $[0.0378, 0.1809]$ across three fits, while step refinement at the selected checkpoint changes energy by about 0.0002. The other three applications retain CFMI's final checkpoint, providing the

Table F.17: Flow validation. (a) Nominal-90% marginal coverage, averaged over targets. (b) Endpoint-selected minus FM-loss-selected joint energy with paired 95% temporal-group intervals; positive favors FM loss. Both use weighted one-standard-error selection on the same trajectory. The other nine model–application combinations all have zero displayed gains and interval endpoints at four decimals. (c) Terminal-minus-selected joint energy across three fits; parentheses give fit SD, and the last column counts positive temporal-group intervals. Refits share the evaluation allocation.

(a) Coverage after checkpoint selection

Application	Final original	Selected original	Selected Direct	Selected SFLA
Gas Turbine	12.2	90.7	66.5	86.6
Appliances	7.3	50.1	67.3	55.8
Beijing	4.9	84.5	75.5	83.2
Household	10.1	91.5	84.7	87.3
Metro	24.5	79.3	85.7	83.0

(b) Endpoint versus flow-loss selection

Application	Flow	Endpoint-selected − FM-loss-selected energy
Gas Turbine	DT-FM	$+0.0051$ $[-0.0248, +0.0350]$
Other nine combinations	DT-FM/CFMI	0.0000 $[0.0000, 0.0000]$

(c) Variation across three refits

Application	Family	Selection gain: mean (fit SD)	Intervals above zero
Beijing	DT-FM	0.2076 (0.0075)	3/3
Beijing	CFMI	0.2082 (0.0058)	3/3
Household	DT-FM	0.3072 (0.0105)	3/3
Household	CFMI	0.2846 (0.0278)	3/3
Metro	DT-FM	0.0791 (0.0059)	3/3
Metro	CFMI	0.0772 (0.0087)	3/3

final-checkpoint/400-step control. DiffPuter retains its 50-step conditional sampler and compares training-loss-best with endpoint-selected epochs.

The statistical controls condition a fitted joint Gaussian mixture or Gaussian copula analytically. Stochastic MICE supplies joint samples, while deterministic MICE and HyperImpute supply point predictions. MIWAE's joint and adaptive importance-resampling procedures are evaluated separately because they yield different conditional predictions. All scores use a common source-fitted coordinate scale.

DiffPuter is likewise evaluated here as a complete-source conditional sampler. Its native incomplete-source EM procedure has a different training protocol.

Table F.18: Completion scores: mean (fit SD), with smaller/larger missing blocks k_s/k_l specified by dataset. Bold marks the lowest displayed mean among procedures completing all three fits, including ties; coverage is unranked. Fits counts eligible source fits shared by both queries. Dashes denote rejected GMM8 fits. Point-only methods: Table F.19.

Procedure	Energy/$\sqrt{k}$ ↓		Mean RMSE ↓		90% coverage		Fits
	k_s	k_l	k_s	k_l	k_s	k_l	
California: $k_s = 2, k_l = 4$							
DT-FM, final	**0.2497** (0.0162)	0.3526 (0.0109)	0.5203 (0.0301)	0.6293 (0.0293)	0.8477 (0.0085)	0.8756 (0.0032)	3/3
DT-FM, selected	**0.2497** (0.0162)	0.3526 (0.0109)	0.5203 (0.0301)	0.6293 (0.0293)	0.8477 (0.0085)	0.8756 (0.0032)	3/3
CFMI, original	0.2580 (0.0137)	0.3576 (0.0098)	0.5366 (0.0182)	0.6400 (0.0302)	0.8548 (0.0082)	0.8693 (0.0044)	3/3
CFMI, selected	0.2580 (0.0136)	0.3576 (0.0098)	0.5368 (0.0182)	0.6401 (0.0300)	0.8636 (0.0073)	0.8744 (0.0032)	3/3
DiffPuter, original	0.2758 (0.0201)	0.3720 (0.0098)	0.5733 (0.0448)	0.7938 (0.2595)	0.8415 (0.0069)	0.8493 (0.0057)	3/3
DiffPuter, selected	0.2821 (0.0240)	0.3800 (0.0105)	3.8994 (4.2826)	6.5836 (6.0097)	0.8333 (0.0122)	0.8457 (0.0047)	3/3
Copula	0.3371 (0.0127)	0.4183 (0.0137)	0.6605 (0.0093)	0.7204 (0.0452)	0.8893 (0.0110)	0.8796 (0.0044)	3/3
GMM8	—	—	—	—	—	—	0/3
Stochastic MICE	0.3331 (0.0179)	0.4283 (0.0105)	0.6147 (0.0367)	0.6979 (0.0292)	0.9128 (0.0104)	0.9163 (0.0064)	3/3
MIWAE, adaptive SIR	0.2523 (0.0146)	**0.3481** (0.0100)	**0.5199** (0.0158)	**0.6177** (0.0340)	0.8496 (0.0045)	0.8587 (0.0012)	3/3
MIWAE, joint SIR	0.3536 (0.0294)	0.4732 (0.0110)	0.6315 (0.0468)	0.7162 (0.0279)	0.5856 (0.0020)	0.5156 (0.0167)	3/3

Continued on next page

Completion scores (continued)

Procedure	Energy/$\sqrt{k}\downarrow$		Mean RMSE $\downarrow$		90% coverage		Fits
	k_s	k_l	k_s	k_l	k_s	k_l	
Protein: $k_s = 2, k_l = 4$							
DT-FM, final	**0.1154** (0.0010)	**0.2000** (0.0013)	**0.2649** (0.0038)	**0.5166** (0.0016)	0.8978 (0.0025)	0.8882 (0.0034)	3/3
DT-FM, selected	**0.1154** (0.0010)	**0.2000** (0.0013)	**0.2649** (0.0038)	**0.5166** (0.0016)	0.8978 (0.0025)	0.8882 (0.0034)	3/3
CFMI, original	0.1186 (0.0009)	0.2038 (0.0024)	0.2733 (0.0015)	0.5224 (0.0041)	0.8903 (0.0025)	0.8792 (0.0047)	3/3
CFMI, selected	0.1186 (0.0008)	0.2038 (0.0024)	0.2734 (0.0015)	0.5223 (0.0042)	0.9020 (0.0015)	0.8854 (0.0020)	3/3
DiffPuter, original	0.1326 (0.0049)	0.2224 (0.0052)	0.5338 (0.2310)	0.7514 (0.3097)	0.8018 (0.0144)	0.8034 (0.0135)	3/3
DiffPuter, selected	0.1324 (0.0046)	0.2213 (0.0035)	0.4817 (0.1734)	0.6295 (0.1014)	0.8021 (0.0142)	0.8115 (0.0123)	3/3
Copula	0.1974 (0.0008)	0.2661 (0.0005)	0.3835 (0.0166)	0.5863 (0.0008)	0.9639 (0.0010)	0.9570 (0.0030)	3/3
GMM8	0.1344 (0.0021)	0.2164 (0.0040)	0.2999 (0.0105)	0.5439 (0.0074)	0.9028 (0.0062)	0.8877 (0.0083)	2/3
Stochastic MICE	0.1840 (0.0044)	0.2601 (0.0040)	0.3356 (0.0026)	0.5639 (0.0019)	0.9124 (0.0031)	0.9056 (0.0029)	3/3
MIWAE, adaptive SIR	0.1187 (0.0022)	0.2070 (0.0022)	0.2677 (0.0097)	0.5346 (0.0027)	0.9082 (0.0110)	0.8799 (0.0018)	3/3
MIWAE, joint SIR	0.3476 (0.0088)	0.4860 (0.0060)	0.5425 (0.0140)	0.7934 (0.0089)	0.4560 (0.0107)	0.3548 (0.0056)	3/3
SGEMM: $k_s = 1, k_l = 2$							
DT-FM, final	0.00230 (0.00018)	**0.00246** (0.00006)	0.00731 (0.00086)	0.00546 (0.00038)	0.9258 (0.0122)	0.9268 (0.00737)	3/3
DT-FM, selected	0.00230 (0.00018)	**0.00246** (0.00006)	0.00731 (0.00086)	0.00546 (0.00038)	0.9258 (0.0122)	0.9268 (0.00737)	3/3
CFMI, original	0.00229 (0.0001)	0.00251 (0.00005)	0.00744 (0.00027)	0.00583 (0.00019)	0.5651 (0.00923)	0.6872 (0.0294)	3/3
CFMI, selected	**0.00221** (0.00008)	**0.00246** (0.00008)	0.00740 (0.00027)	0.00587 (0.0002)	0.9030 (0.0188)	0.9037 (0.0133)	3/3
DiffPuter, original	0.00848 (0.00086)	0.0112 (0.00106)	0.0157 (0.00129)	0.0188 (0.00168)	0.8997 (0.00113)	0.8434 (0.0144)	3/3
DiffPuter, selected	0.00819 (0.00103)	0.0106 (0.00143)	0.0153 (0.00144)	0.0182 (0.00209)	0.9167 (0.0157)	0.8691 (0.0364)	3/3
Copula	0.0627 (0.00106)	0.0738 (0.00133)	0.1413 (0.00257)	0.1486 (0.00334)	1.0000 (0.0000)	1.0000 (0.0000)	3/3
GMM8	0.00381 (0.00008)	0.00423 (0.0001)	**0.00729** (0.00013)	0.00547 (0.00007)	0.9935 (0.00298)	0.9906 (0.00482)	3/3
Stochastic MICE	0.00254 (0.00008)	0.00276 (0.00012)	0.00758 (0.00006)	**0.00499** (0.00006)	0.9622 (0.0118)	0.9587 (0.00842)	3/3
MIWAE, adaptive SIR	0.00281 (0.00019)	0.00301 (0.00017)	0.00835 (0.00025)	0.00608 (0.00019)	0.9603 (0.00407)	0.9603 (0.0121)	3/3
MIWAE, joint SIR	0.0584 (0.00971)	0.1993 (0.00747)	0.1570 (0.0356)	0.4007 (0.0265)	0.2988 (0.0269)	0.1582 (0.0208)	3/3
Tetouan: $k_s = 2, k_l = 4$							
DT-FM, final	0.4666 (0.0314)	0.5089 (0.0204)	0.7180 (0.0353)	0.6821 (0.0232)	0.2952 (0.0226)	0.3304 (0.0071)	3/3
DT-FM, selected	0.3870 (0.0285)	**0.4095** (0.0273)	0.6592 (0.0422)	**0.6375** (0.0354)	0.7282 (0.0121)	0.7677 (0.0175)	3/3
CFMI, original	0.4863 (0.0394)	0.5296 (0.0266)	0.7360 (0.0437)	0.6998 (0.0304)	0.2783 (0.0181)	0.2957 (0.0045)	3/3
CFMI, selected	**0.3838** (0.0375)	0.4114 (0.0290)	**0.6590** (0.0572)	0.6421 (0.0348)	0.7852 (0.0468)	0.8031 (0.0548)	3/3
DiffPuter, original	0.4283 (0.0231)	0.4906 (0.0176)	0.6709 (0.0347)	0.6643 (0.0290)	0.2366 (0.0264)	0.2651 (0.0216)	3/3
DiffPuter, selected	0.4313 (0.0298)	0.4702 (0.0305)	0.6848 (0.0447)	0.6567 (0.0329)	0.2852 (0.0575)	0.3337 (0.0594)	3/3
Copula	0.3968 (0.0226)	0.4324 (0.0216)	0.7150 (0.0354)	0.6946 (0.0327)	0.8604 (0.0110)	0.8737 (0.0103)	3/3
GMM8	0.4747 (0.0580)	0.5044 (0.0753)	0.7883 (0.0784)	0.7768 (0.1000)	0.5977 (0.0424)	0.6514 (0.0351)	3/3
Stochastic MICE	0.3993 (0.0070)	0.4166 (0.0142)	0.6750 (0.0116)	0.6467 (0.0205)	0.7992 (0.0209)	0.8328 (0.0105)	3/3
MIWAE, adaptive SIR	0.4199 (0.0117)	0.4581 (0.0238)	0.6804 (0.0240)	0.6541 (0.0487)	0.4251 (0.0232)	0.4277 (0.0461)	3/3
MIWAE, joint SIR	0.4732 (0.0247)	0.5332 (0.0363)	0.7094 (0.0400)	0.6959 (0.0527)	0.2855 (0.0197)	0.2765 (0.0194)	3/3

Extreme draws and predictive means. On California, selected DiffPuter has mean RMSE $3.8994/6.5836$ for the two query sizes, versus $0.5733/0.7938$ for the original. Two of three fits produce extreme draws: five rows account for over 95% of squared mean error, while median RMSE is about 0.608. Extreme draws also appear in source validation.

The discrepancy is possible because energy does not control squared mean error. Against truth zero, $Q = (1-\epsilon)\delta_0 + \epsilon\delta_M$ has energy $\epsilon^2|M|$ and mean error $\epsilon|M|$. Taking $M = \epsilon^{-3/2}$ makes the former vanish and the latter diverge. We therefore retain point error and coverage alongside energy, using the same predictions without metric-specific reselection.

On Tetouan, selection raises nominal-90% coverage from 29.5% to 72.8% for DT-FM and 27.8% to 78.5% for CFMI. Both still undercover. A fixed-model SGEMM check raises CFMI coverage from about 63% to 91% at 100 versus 400 Euler steps with only about 3% energy change, showing why coverage is reported separately from average score.

Point prediction. ReMasker (Du et al., 2024) and PMAE (Kim et al., 2025) each use three 600-epoch terminal fits on the same source rows, query masks and scale. PMAE's fitting-only artificial-missingness configuration transfers from California to the other tables. Their outputs are evaluated as point predictions under their specified training budgets.

Training-context coverage. Spreading the same 4,176 Tetouan fitting rows over 70–75 rather than 29 days improves energy in all three paired fits for both selected flows and the copula. Mean gains are 12.7–12.8%, 8.8–10.0% and 5.2–5.4%, respectively. Coverage remains below nominal. Because temporal dependence and operating-condition coverage change together, this diagnoses sensitivity to source allocation. The main benchmark retains the original allocation.

Table F.19: Matched point accuracy: RMSE at the smaller/larger query size, with standard deviations across three fits in parentheses. Generative methods contribute their predictive means; ReMasker, PMAE, HyperImpute and MICE contribute point outputs. The methods use their specified training schedules and different compute budgets.

Procedure Missing coordinates	California 2/4	Protein 2/4	SGEMM 1/2	Tetouan 2/4
DT-FM, selected	0.5203/0.6293 (0.0301/0.0293)	**0.2649/0.5166** (0.0038/0.0016)	**0.0073**/0.0055 (0.0009/0.0004)	0.6592/**0.6375** (0.0422/0.0354)
CFMI, selected	0.5368/0.6401 (0.0182/0.0300)	0.2734/0.5223 (0.0015/0.0042)	0.0074/0.0059 (0.0003/0.0002)	**0.6590**/0.6421 (0.0571/0.0348)
DiffPuter, original	0.5733/0.7938 (0.0449/0.2595)	0.5338/0.7514 (0.2310/0.3097)	0.0157/0.0188 (0.0013/0.0017)	0.6709/0.6643 (0.0347/0.0290)
MIWAE, adaptive SIR	**0.5199/0.6177** (0.0157/0.0340)	0.2677/0.5345 (0.0097/0.0027)	0.0083/0.0061 (0.0003/0.0002)	0.6804/0.6541 (0.0240/0.0487)
MICE, point	0.6091/0.6876 (0.0346/0.0283)	0.3313/0.5610 (0.0023/0.0006)	0.0075/**0.0049** (0.0001/0.0000)	0.6720/0.6435 (0.0111/0.0227)
PMAE	0.5501/0.6486 (0.0258/0.0161)	0.2663/0.5169 (0.0154/0.0029)	0.0926/0.0869 (0.0125/0.0123)	0.6663/0.6703 (0.0520/0.0324)
ReMasker	0.6251/0.6267 (0.1228/0.0339)	0.3435/0.5270 (0.0228/0.0064)	0.1359/0.1246 (0.0277/0.0289)	0.6614/0.6702 (0.0386/0.0141)
HyperImpute	0.5695/0.7802 (0.0229/0.0425)	0.3662/0.6192 (0.1381/0.0449)	0.00762/0.00577 (0.00011/0.00003)	0.6345/0.6578 (0.0576/0.0627)

F.5 Joint prediction from selectively recorded labels

We now make the training responses selectively observed as well. All procedures share the four tables, data split, weights and response-access restrictions, extending the complete-source benchmark in Appendix F.4 to incomplete training data.

Four-table comparison. California, Protein, SGEMM and Tetouan each provide a task predicting the first two eligible measured columns. A separate non-target measurement defines the recording state but is excluded from all inputs. Its above-median indicator gives the informative condition, while an independent fair bit gives the unrelated-state control. Target recording probabilities are 0.3 and 0.7 in both conditions. In the null the state is unrelated to the response, so transport opportunity vanishes although fitted updates can still act.

The source contains 4,352 rows, or 4,432 rows across 111 days for Tetouan. Groups are assigned to fitting, mixture estimation, checkpoint selection and procedure selection in proportions $60/20/10/10$. Only recorded source responses supply labels. Evaluation starts from 512 separate rows and scores the unrecorded blocks. Ridge logistic state probabilities and smoothed recording probabilities supply contextual source/deployment mixtures. Fitting uses scales from observed fitting-set values. Reporting uses the fixed benchmark scale. All methods share the source allocation, target geometry and evaluation rows. Each table has one inspected allocation, with no independent confirmation from adding model families.

ReMasker and PMAE use native training configurations at their final checkpoints. HyperImpute (Jarrett et al., 2022) uses its cycle-average adapter. The chained-equation approach of MICE (van Buuren & Groothuis-Oudshoorn, 2011) contributes its default point prediction, a 100-sweep point version and a separately named stochastic sampler with fixed fitted conditionals. Native methods may use partially observed rows, whereas the conditional flows fit eligible recorded target blocks. The flow comparison here retains its source-selected checkpoints for both DT-FM and CFMI. Training, stopping and computation therefore differ across families. The comparison evaluates the complete procedures under their specified training and prediction rules. Native incomplete-source DiffPuter EM performance remains unevaluated.

Competitiveness and correction. Both original flow predictors improve energy over stochastic MICE on California and Protein. CFMI also improves on SGEMM (Table F.21). Tetouan's intervals include zero and favor MICE in their means. Point prediction gives a different ordering: SGEMM's MICE and HyperImpute means are more accurate than the flow means. The native neural imputers and statistical controls therefore remain useful comparators even where flows give stronger distributional scores.

Table F.20: Point prediction from incomplete source labels: RMSE ↓. Each application compares original predictions and SFLA on the same labels. For point methods, SFLA averages conditional means to produce a point prediction. The control keeps unequal recording rates but makes the recording state unrelated to the response.

(a) Informative recording

	California		Protein		SGEMM		Tetouan	
Predictor	Original	SFLA	Original	SFLA	Original	SFLA	Original	SFLA
DT-FM	0.7483	**0.7408**	0.0565	0.0723	0.0083	0.0089	0.6793	0.6796
CFMI	0.7690	0.7541	0.0596	**0.0541**	0.0075	0.0070	0.6597	0.6569
Gaussian	0.9005	0.8783	0.1451	0.1060	0.0078	0.0219	0.6686	**0.6401**
GMM4	1.1069	1.0309	0.0869	0.0847	0.0115	0.0264	0.7073	0.8098
Copula	0.8154	0.8008	0.6560	0.1019	0.3122	0.2024	0.6760	0.6535
MICE point	0.8909	0.8936	0.1428	0.1424	**0.0051**	**0.0051**	0.6677	0.6690
MICE, 100 sweeps	0.8909	0.8936	0.1427	0.1423	**0.0051**	**0.0051**	0.6677	0.6691
Stochastic MICE	0.8989	0.9017	0.1438	0.1434	0.0055	0.0055	0.6747	0.6762
HyperImpute	0.7634	0.7415	0.1554	0.1563	0.0052	0.0052	0.6983	0.6978
ReMasker	0.8118	0.8875	0.3825	0.2654	0.1748	0.1243	0.7412	0.7551
PMAE	0.8831	0.8806	0.1112	0.1302	0.0064	0.0060	0.7006	0.7309

(b) Independent-state control

	California		Protein		SGEMM		Tetouan	
Predictor	Original	SFLA	Original	SFLA	Original	SFLA	Original	SFLA
DT-FM	0.7483	0.7543	**0.0476**	0.0540	0.0110	0.0112	0.7809	0.7873
CFMI	0.7470	0.7259	0.0484	0.0478	0.0086	0.0100	0.7859	0.7823
Gaussian	0.7860	0.7887	0.1299	0.1297	0.0081	0.0084	0.7189	0.7160
GMM4	0.7920	0.7598	0.0774	0.0785	0.0104	0.0115	0.7164	0.7180
Copula	0.7620	0.7643	0.1927	0.2047	0.5433	0.6785	0.7325	0.7317
MICE point	0.7860	0.7868	0.1293	0.1293	**0.0059**	**0.0059**	0.7111	**0.7066**
MICE, 100 sweeps	0.7860	0.7868	0.1293	0.1293	**0.0059**	**0.0059**	0.7111	**0.7066**
Stochastic MICE	0.7904	0.7912	0.1306	0.1306	0.0060	0.0060	0.7135	0.7088
HyperImpute	0.7099	**0.6944**	0.0974	0.0968	0.0060	0.0061	0.8213	0.7818
ReMasker	0.7843	0.7609	0.2461	0.1900	0.2164	0.1901	0.7268	0.7756
PMAE	0.8600	0.8036	0.0506	0.0580	0.0076	0.0070	0.7759	0.8112

Dependence and population updating. Completed-pair covariance tests a different use of the predictions from RMSE or marginal coverage. Table F.22 retains source assembly to distinguish building the source mixture from changing its probabilities. On Tetouan, CFMI's error falls from 0.0420 to 0.0051 on constructing source assembly, then rises to 0.0068 after updating. ReMasker's Protein error instead rises from 0.0637 to 0.2229 when building the source mixture, with little further change after updating. Neither original-to-deployment contrast alone measures the effect of recording correction.

DT-FM has low covariance error on Protein, and CFMI does on Tetouan. ReMasker gives the lowest displayed California error. On SGEMM, ReMasker's deployment error is 0.0722, while several other methods have small errors. Both flow errors increase after updating on California.

For point methods, assembly averages state-conditional means and returns a point estimate. Writing $m_j(H) = E[Y_j \mid H]$,

$$\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}\{m_1(H), m_2(H)\} + E\{\mathrm{Cov}(Y_1, Y_2 \mid H)\}.$$

Variation between predicted means can therefore recover some completed-row covariance even when conditional spread is absent. Covariance error, proper scores and conditional-spread checks measure different properties of the predictions. Figure 3 isolates the dependence lost by replacing a distribution with its mean or factorized marginals, and the measured-data comparison assesses the practical consequences.

Sampling a state and returning its conditional mean would preserve between-state variation but omit within-state residual variation. The point-averaging procedures evaluated here have no such sampling step.

Calibration remains limited after assembly: nominal 90% marginal coverage is about 80%/85% for DT-FM/CFMI on California, 83%/77% on Protein, 67%/89% on SGEMM, and 65%/72% on Tetouan. Results with a state unrelated to the response are retained in Tables F.20 and F.21.

Table F.21: Distributional prediction from incomplete source labels. Energy is divided by $\sqrt{2}$ and uses 64 draws (↓). Positive gains favor the original flow over stochastic MICE; paired 95% group intervals condition on the fitted models. The two recording conditions share the same comparison rules.

(a) Informative recording

	California		Protein		SGEMM		Tetouan	
Predictor	Original	SFLA	Original	SFLA	Original	SFLA	Original	SFLA
DT-FM	**0.4574**	0.4578	0.0296	0.0311	0.0039	0.0045	0.4357	0.4348
CFMI	0.4735	0.4645	0.0286	**0.0283**	0.0028	**0.0025**	0.4185	0.4117
Gaussian	0.5283	0.5196	0.0789	0.0577	0.0126	0.0149	0.4017	**0.3846**
GMM4	0.5435	0.5291	0.0452	0.0411	0.0127	0.0162	0.4325	0.4835
Copula	0.4982	0.4897	0.2086	0.0629	0.1070	0.0546	0.4063	0.3951
Stochastic MICE	0.5292	0.5305	0.0786	0.0777	0.0042	0.0042	0.4049	0.4059

(b) Independent-state control

	California		Protein		SGEMM		Tetouan	
Predictor	Original	SFLA	Original	SFLA	Original	SFLA	Original	SFLA
DT-FM	0.4531	0.4595	0.0251	0.0275	0.0046	0.0049	0.5085	0.5168
CFMI	0.4523	**0.4420**	0.0253	**0.0248**	**0.0031**	0.0035	0.4908	0.4934
Gaussian	0.4770	0.4792	0.0713	0.0714	0.0116	0.0117	0.4320	0.4299
GMM4	0.4793	0.4637	0.0417	0.0416	0.0119	0.0121	0.4428	0.4412
Copula	0.4618	0.4630	0.1121	0.1114	0.1683	0.1800	0.4377	0.4374
Stochastic MICE	0.4837	0.4843	0.0712	0.0712	0.0034	0.0034	0.4285	**0.4253**

(c) Original-flow gains over stochastic MICE

Table	Flow	Rows	Groups	Energy gain over stochastic MICE [95% interval]
California	DT-FM	262	262	**0.0718** [0.0304,0.1132]
California	CFMI	262	262	**0.0557** [0.0147,0.0967]
Protein	DT-FM	248	248	**0.0490** [0.0401,0.0579]
Protein	CFMI	248	248	**0.0500** [0.0411,0.0588]
SGEMM	DT-FM	264	264	0.0004 [-0.0003,0.0010]
SGEMM	CFMI	264	264	**0.0014** [0.0008,0.0020]
Tetouan	DT-FM	247	54	-0.0308 [-0.0865,0.0248]
Tetouan	CFMI	247	54	-0.0136 [-0.0702,0.0431]

Selection with a matched source library. The source library contains original, Direct, source assembly and their half-original/half-source-assembly blend. Adding deployment assembly and its half blend leaves all sixteen selected predictions unchanged: fourteen originals and two source blends. Selection requires positive source evidence after accounting for multiple comparisons and recording-probability uncertainty. Table F.23 measures its conservatism: the original can be retained even when another candidate has lower evaluation risk. The largest retrospective gap is 0.01627. Each table supplies one shared allocation, so the model/condition rows are dependent.

The evaluation minima diagnose missed opportunities within the library. They play no part in source-based selection.

Attribution within a selected blend. A separate five-family California comparison selects the half-original, half-SFLA blend for DT-FM under informative recording and the original in the other nine cells. Its joint energy changes from 0.4574 (original) to 0.4467 (source-assembly blend) and 0.4460 (deployment-assembly blend). The total gain is 0.0114 $[0.0025, 0.0203]$, while the matched updating increment is only 0.0007 $[-0.0022, 0.0036]$. Most of this improvement comes from fitting and blending the source components. In the four-table comparison above, both blends are available and adding deployment assembly leaves the selected predictions unchanged.

F.6 Adult income

The classification comparison uses two overlapping recording assignments, each with informative and outcome-unrelated states. LoCalPFN uses 1,024 fitting labels, 512 validation labels, retrieval and 630 fine-tuning updates. Native AUC and weighted Brier selection return the same checkpoint. Expanded

Table F.22: Absolute completed-pair covariance error ↓ under informative recording. Source assembly and SFLA use source and deployment state weights. The top six rows return samples; the bottom five average point predictions. Bold marks the lowest displayed error per application. Values are descriptive, from one allocation per table on fixed benchmark scales.

	California			Protein		
Predictor	Original	Source assembly	SFLA	Original	Source assembly	SFLA
DT-FM	0.0736	0.0800	0.0908	**0.0045**	0.0089	0.0116
CFMI	0.0649	0.0765	0.0865	0.0192	0.0213	0.0182
Gaussian	0.0736	0.0601	0.0646	0.0490	0.0242	0.0177
GMM4	0.3345	0.1344	0.1050	0.0211	0.0276	0.0269
Copula	0.1139	0.0920	0.0971	1.0498	0.0092	0.0116
MICE, stochastic	0.0761	0.0702	0.0721	0.0505	0.0452	0.0408
ReMasker	0.0443	0.0233	**0.0194**	0.0637	0.2229	0.2245
PMAE	0.1109	0.0722	0.0765	0.0617	0.1034	0.1007
HyperImpute	0.1041	0.1234	0.1256	0.1135	0.1091	0.1061
MICE, point	0.0801	0.0737	0.0779	0.0539	0.0485	0.0441
MICE, point (100 sweeps)	0.0801	0.0737	0.0779	0.0531	0.0477	0.0433

	SGEMM			Tetouan		
Predictor	Original	Source assembly	SFLA	Original	Source assembly	SFLA
DT-FM	0.0007	**0.0003**	0.0004	0.0457	0.0520	0.0438
CFMI	0.0006	0.0006	0.0007	0.0420	**0.0051**	0.0068
Gaussian	0.0013	0.0008	0.0013	0.2515	0.1395	0.1078
GMM4	0.0012	0.0008	0.0019	0.1570	0.3075	0.2534
Copula	0.3832	0.0589	0.0841	0.1755	0.1350	0.1186
MICE, stochastic	0.0006	0.0007	0.0007	0.2508	0.2452	0.2465
ReMasker	0.0093	0.0760	0.0722	0.2532	0.2715	0.2714
PMAE	0.0027	0.0025	0.0027	0.1418	0.2220	0.2222
HyperImpute	0.0009	0.0009	0.0009	0.1175	0.1121	0.1143
MICE, point	0.0004	0.0004	0.0005	0.0980	0.0925	0.0939
MICE, point (100 sweeps)	0.0005	0.0005	0.0005	0.0983	0.0927	0.0942

Table F.23: Prediction gain and selection error for the incomplete-source flows. Gain is original minus returned energy. Selection gap is returned minus the best evaluated energy in the expanded library. The source-only and expanded libraries return the same predictor in every case. Evaluation minima are retrospective diagnostics. All rows share one allocation per table.

		DT-FM		CFMI	
Application	Recording	Gain ↑	Selection gap ↓	Gain ↑	Selection gap ↓
California	Informative	0.0107	0.00069	0.0000	0.0147
California	Unrelated state	0.0000	0.0163	0.0000	0.0117
Protein	Informative	0.0000	0.00053	0.0000	0.00131
Protein	Unrelated state	0.0000	0.00003	0.0000	0.00121
SGEMM	Informative	0.0002	0.00011	0.0000	0.00024
SGEMM	Unrelated state	0.0000	0.00036	0.0000	0.0000
Tetouan	Informative	0.0000	0.00609	0.0000	0.00849
Tetouan	Unrelated state	0.0000	0.0000	0.0000	0.00451

TabICLv2 uses more labels. Table F.24 retains the native/retrieval references, fixed correction routes and the two published marginal corrections under both assignments and conditions.

Expanded TabICLv2 SFLA beats matched LoCalPFN in all four conditions, with paired gaps .0209–.0336 and intervals excluding zero. Its own original is better in three conditions, so much of this advantage precedes correction. Both fitted heads return zero updates. TabDPT Context Direct loses in all four conditions. Its favorable shifted SFLA difference has an interval containing zero.

DistPFN and DistPFN-T use saved class probabilities, source class frequencies and the unlabelled query batch. Their shifted gains over the original have intervals containing zero in every case. Both worsen the TabICLv2 and TabDPT shifted mean risks. LoCalPFN's temperature variant improves its two shifted point estimates. Paired intervals use 1,000 duplicate-group bootstrap resamples, recomputing batch correction and holding classifiers and source priors fixed.

Table F.24: Adult Brier risk ↓. A1/A2 are overlapping assignments. "Unrelated" means that recording information carries no additional outcome information. TabICLv2 rows use expanded contexts, with a larger label budget than matched LoCalPFN. Bold marks lowest displayed means. Source assembly and SFLA hold the component predictions fixed.

Prediction	A1 unrelated	A1 shift	A2 unrelated	A2 shift
TabPFN: native 1000	0.2231	0.2343	0.2108	0.2139
LoCalPFN: AUC-selected	0.2234	0.2261	0.2082	0.2140
LoCalPFN: context Direct	0.2327	0.2206	0.2075	0.2121
TabICLv2: original, expanded contexts	0.2015	0.1985	0.1832	0.1791
TabICLv2: source assembly	0.2025	0.1992	0.1849	0.1813
TabICLv2: deployment assembly	0.2025	0.1977	0.1849	0.1803
TabICLv2: context Direct	0.2045	0.2057	0.1856	0.1896
TabICLv2: DistPFN	0.2015	0.2029	0.1832	0.1869
TabICLv2: DistPFN-T	0.2060	0.1989	0.1841	0.1798
TabDPT: original	**0.1975**	0.1966	**0.1735**	**0.1772**
TabDPT: uniform context	0.2030	0.2041	0.1853	0.1853
TabDPT: Direct	0.2030	0.2007	0.1853	0.1950
TabDPT: source assembly	0.1979	0.1971	0.1800	0.1801
TabDPT: SFLA	0.1979	**0.1959**	0.1800	0.1786
TabDPT: DistPFN	0.1977	0.2013	0.1739	0.1866
TabDPT: DistPFN-T	0.2022	0.1972	0.1741	0.1787
LoCalPFN: DistPFN	0.2235	0.2263	0.2095	0.2127
LoCalPFN: DistPFN-T	0.2246	0.2246	0.2082	0.2094

F.7 Clinical laboratory values

The clinical example shows a useful mixture update whose net prediction gain remains uncertain. We use 3,453 eligible stays from the PhysioNet/Computing in Cardiology Challenge 2012, version 1.0.0. Targets are the latest valid nonnegative hematocrit and platelet measurements before hour six, with timestamp ties resolved by recorded order. Inputs contain a visibility bit and, when visible, Glasgow Coma Scale categories below 8, 8–14 and at least 15. This gives four distinct inference inputs.

Recording is imposed after splitting, retaining ground truth for evaluation. Within-split response ranks combine independent and comonotone assignments with weights $.25/.75$; the control uses independence. In $(h, C) = (0,0), (0,1), (1,0), (1,1)$ order, masses are proportional to $(59, 148, 958, 1235)$ and recording probabilities are $(18/59, 83/148, 548/958, 948/1235)$. Only recorded outcomes train the models, and C is omitted at inference. This evaluates imposed recording on clinical measurements; natural missingness would require a justified bridge.

Four assignment-and-refit procedures share the stays, including a 689-stay evaluation set. Missing-outcome counts are 223–243 under shift and 224–247 under the control. Refit-t intervals describe assignment and fitting variation. The 1,999 shared-stay bootstrap resamples retain dependence across refits. Flow and statistical controls have the same label access. Source validation uses 128 draws and evaluation 1,024. Copula weighting applies to marginals and dependence. Configuration and licence are in Table C.3, C.2.

Updating improves all four shifted flow refits, with mean energy gain 0.014447, offsetting a 0.011648 loss from fitting the source mixture. The total is 0.002800 $[-0.000784, 0.006383]$. Direct's changes have both signs for flows, while both correction routes improve the simple distributional models in every shifted refit.

Nominal-90% marginal coverage is $.899/.894$ for assembly and $.890/.879$ for Direct. Assembly assigns $.00132$ mass to negative values. Gains for the event that both outcomes exceed their source-fitting 75th percentiles remain uncertain. Selecting among flow, simple and combined libraries also leaves failures: simple-only selection harms its reference in two shifted refits and combined selection in one. Every policy class harms prediction in one control refit. Thus improved mixture energy does not establish improvement for every functional or selected procedure.

Table F.25: Clinical normalized energy and attribution. (a) Means over four assignment-and-refit procedures. Both assemblies omit the individual state at inference. (b) Refit intervals use four procedure means; shared-stay intervals retain dependence across refits when resampling evaluation stays.

(a) Predictive accuracy across model families

Model	Original	Weighted Direct	Source assembly	SFLA
Conditional flow	0.6131	0.6134	0.6247	**0.6103**
Gaussian	0.6299	0.6146	0.6276	0.6130
Student-t	0.6298	0.6153	0.6273	0.6149
GMM4	0.6283	0.6134	0.6330	0.6203
Weighted copula	0.6218	0.6113	0.6226	0.6119

(b) Paired flow gains

Condition	Flow correction	Mean gain	Refit interval	Shared-stay interval
Null	Direct weighting	-0.002755	[-0.0158,0.0103]	[-0.0076,0.0021]
Null	Mixture update	0.000160	[-0.0015,0.0019]	[-0.0003,0.0006]
Positive	Direct weighting	-0.000278	[-0.0153,0.0147]	[-0.0048,0.0045]
Positive	Mixture update	0.014447	[0.0062,0.0227]	[0.0048,0.0237]

G Related work and scope

Our distributional procedures build on identification under selective observation. Their connections to repeated regression, imputation and model adaptation clarify which parts of the approach are inherited and where the implemented outputs differ. We also develop the potential-outcome interpretation and state the limits of the empirical evidence.

G.1 Identification and related estimators

The closest setting is privilegedly missing at random. Boeken et al. (2023) identify the conditional distribution using information available during estimation but absent at inference. They develop repeated regression, importance weighting and doubly robust regression for the conditional mean. Their repeated-regression estimator fits a mean using (H, C), then regresses its pseudo-responses on H using auxiliary records. The published comparisons use splines and regression trees under squared error. SFLA instead combines the state-conditional distributions directly, retaining their variation and dependence without an outer outcome regression.

For a single binary outcome, the mean defines a Bernoulli distribution, so both implemented estimators provide probabilistic correction. With shared tree components on continuous outcomes, neither generally dominates in RMSE (Appendices D.1.2 and F.1.1). Appendix B.11 discusses an extension that learns from teacher distributions or joint samples. It goes beyond the mean-regression implementation evaluated by Boeken et al. (2023).

Direct uses the change-of-measure identity underlying importance weighting and target-oriented validation (Shimodaira, 2000; Sugiyama et al., 2007), applied on (H, C) while prediction uses H. Auxiliary variables also improve imputation outside the final analysis model (Collins et al., 2001). Shannon et al. (2026) formulate imputation as weighted risk minimization under MAR. The appropriate weights depend on whether prediction targets unrecorded cases or all eligible cases (Appendix A.4). Pattern-mixture models (Little, 1993) and mixtures of experts (Jacobs et al., 1991) provide related representations. SFLA holds the conditional distributions fixed and changes their mixing probabilities. Its fitted-error analysis separates this update from fitting the components, whose errors can also make prediction-optimal weights differ from population proportions (Appendix B.11).

Selective-label methods use several identification strategies. Verification-bias correction (Begg & Greenes, 1983) adjusts selective outcome ascertainment. Decision-maker variation (Lakkaraju et al., 2017) and instrumental variables (Chen et al., 2025) can supply other assumptions. Our bridge relies on observed information sufficient to adjust for recording. Selection can also distort fairness assessments (Kallus & Zhou, 2018; Coston et al., 2021; Fawkes et al., 2024). DCEM (Chang &

Wiens, 2024) specifically models group-dependent testing with negative proxy labels for untested cases. Our outcomes are verified labels. Appendix A.5 works through a credit-score audit, including the additional group-specific bridge and counterfactual outcome requirements. The paper evaluates predictive correction rather than a fairness intervention.

G.2 Predictive models and adaptation

Masks and the prediction target. ReMasker (Du et al., 2024) and PMAE (Kim et al., 2025) reconstruct artificially withheld values. MaskTab (Zheng et al., 2026) and CFMI (Šimkus & Gutmann, 2025) distinguish recording and artificial masks, and CFMI states an observed-to-missing generalization assumption. Separate masks and informative recording thus have precedents. Our correction addresses a population difference that retaining masks alone can leave unresolved (Appendix A.2). Missingness changes need not change the Bayes predictor (Rockenschaub et al., 2024). Adaptation is needed here when changing $P(C \mid H)$ changes $P(Y \mid H)$ despite stability at (H, C).

Generative imputers estimate distributions through different objectives: conditional diffusion in CSDI (Tashiro et al., 2021), observed-data likelihood in MIWAE (Mattei & Frellsen, 2019), diffusion with EM in DiffPuter (Zhang et al., 2025), and averaging over latent completions in Missing-Data Flow Matching (Khan et al., 2026). MIRI (Yu et al., 2025) reduces dependence between completed data and missingness indicators. Recording correction specifies which population their predictions should represent. When the bridge fails, alternatives such as not-MIWAE (Ipsen et al., 2021) require assumptions about the missingness process. Appendix A.2 distinguishes these targets.

Frozen tabular predictors. TabPFN (Hollmann et al., 2025; Jäger et al., 2026) and TabICLv2 (Qu et al., 2026) pretrain on synthetic tasks, while TabDPT (Ma et al., 2026a) uses real tables. Their labelled contexts retain the source population's selection. LoCalPFN (Thomas et al., 2024) combines retrieval and fine-tuning, and CRUMB (Heredge et al., 2026) matches contexts to query features. Context Direct instead uses recording weights in context sampling. Since this need not implement weighted risk minimization, our uniform-context and matched-learning controls evaluate the two operations separately. Native missing-feature interfaces are described in Appendix C.4.

Label-shift methods adjust class frequencies under stability of features given class (Lipton et al., 2018; Lee, 2026). Our unchanged-marginals example preserves feature, state and label marginals while changing conditional state proportions. It separates the information missing from marginal summaries from the restriction to global class adjustments. Latent-subgroup adaptation (Alabdulmohsin et al., 2023) instead uses concepts and proxies to identify hidden population changes. Here C is observed in estimation records, allowing direct estimation of $P_0(C \mid H)$.

Joint outputs and validation. Flows, mixtures, copulas and models such as Engression (Shen & Meinshausen, 2023) can supply conditional distributions. Proper scores (Gneiting & Raftery, 2007) evaluate their predictions, but energy can miss important dependence errors (Pinson & Tastu, 2013); we therefore also measure associations. Flow matching (Lipman et al., 2023; Liu et al., 2023) learns transport fields by regression. Our checkpoint controls credit validation to both predictive scores and held-out flow loss. Early stopping and inductive bias have established roles in diffusion generalization (Favero et al., 2025; Kadkhodaie et al., 2024), while TabCutMix (Fang et al., 2025) changes training endpoints by exchanging feature groups. They motivate separate controls for training and checkpoint choice. Predictive dispersion alone provides insufficient evidence of record copying. An outer model could also learn from SFLA's joint samples to reduce prediction cost (Appendix B.11). Such distillation would need to preserve the relevant dependence.

G.3 Potential-outcome prediction

The same estimators apply to potential outcomes when causal assumptions identify the conditional distributions (Hernán & Robins, 2020). Treatment selects which potential outcome is observed. Additional information C remains an adjustment variable.

A trial and its intended patient population. Consider blood pressure six months after treatment d. Routine inputs H include age and baseline blood pressure, while C contains a pretreatment frailty score and detailed comorbidity assessment. Trial participants supply (H, C, D, Y). A rep-

resentative target sample supplies (H, C) without requiring treatment or follow-up outcomes. A predictor can thus use assessments collected for a population study without requiring them at every subsequent prediction. Restricting prediction to H is an application choice, additional to standard trial generalization.

Dahabreh et al. (2021) study this combination of trial outcomes and target covariates, with our (H, C) corresponding to their baseline adjustment variables. Their sampling design distinguishes all eligible patients from nonparticipants. Dahabreh et al. (2023) extend this data pattern to several trials, and Robertson et al. (2024) distinguish subgroup-defining variables from adjustment variables. These works target potential-outcome means. Full distributions require the stronger distributional assumptions below. Trial-engagement effects also need care: conditions for transporting average treatment effects (Ung et al., 2025) need not identify each arm's outcome distribution.

Assumptions and estimators. Let D be a discrete treatment, distinct from prediction query A, and $Y(d)$ its potential outcome. Let P_s denote the source and $W = (H, C)$ contain pretreatment covariates. Assume consistency $Y = Y(D)$ for a well-defined intervention, the absence of relevant interference, exchangeability $Y(d) \perp D \mid W$, and positivity $e_d(W) = P_s(D = d \mid W) > 0$ on target support. Then

$$P_s(Y(d) \in dy \mid W) = P_s(Y \in dy \mid W, D = d) =: Q_d(dy \mid W).$$

For a different target population, additionally require $P_0^W \ll P_s^W$ and $P_0(Y(d) \in dy \mid W) = P_s(Y(d) \in dy \mid W)$. Randomization supports within-trial exchangeability, while these latter conditions justify population transfer. Post-treatment measurements cannot be substituted for C without further causal justification. Missing factual outcomes require a separate recording assumption.

For an integrable loss $L_\theta(H, Y)$, define $m_{d,\theta}(W) = \int L_\theta(H, y) Q_d(dy \mid W)$ and $a(W) = dP_0^W / dP_s^W$. Direct uses the weighted risk

$$R_{0,d}(\theta) := \mathbb{E}_0 L_\theta(H, Y(d)) = \mathbb{E}_0 m_{d,\theta}(W) = \mathbb{E}_s\left[\frac{a(W)\mathbf{1}_{\{D=d\}}}{e_d(W)} L_\theta(H, Y)\right]. \tag{G.1}$$

SFLA fits treatment-specific distributions and averages over the target covariates, $\Pi_0(dc \mid h) = P_0(C \in dc \mid H = h)$:

$$P_0(Y(d) \in dy \mid H = h) = \int Q_d(dy \mid h, c)\, \Pi_0(dc \mid h). \tag{G.2}$$

Proof. Exchangeability and consistency identify Q_d. Integrating this law over $P_0(C \mid H)$ proves the assembly identity. For Direct, conditional expectation given W reduces the weighted factual loss to $a(W) m_{d,\theta}(W)$, whose source expectation is the target risk by change of measure. □

Within the same population, $a = 1$. For transfer across populations, the target sample estimates either a or Π_0. Flows and frozen predictors can supply Q_d through a suitable distributional interface. Proper-score learning targets that distribution, while flow matching also depends on field estimation and numerical integration. The weighted-learning identity applies to the objective and gives no guarantee for context resampling alone.

Augmented and causal learners. DoubleGen (Luedtke & Fukumizu, 2025) develops augmented generative learning; GDR-learners (Melnychuk & Feuerriegel, 2026) also allow prediction inputs to omit adjustment variables. DiffPO (Ma et al., 2024) combines propensity-weighted denoising with local orthogonality at a correctly specified conditional optimum. DecFM (Dance et al., 2026) corrects an estimated observational starting distribution.

CausalPFN (Balazadeh et al., 2025), CausalFM (Ma et al., 2026b) and Do-PFN (Robertson et al., 2025) use causal pretraining to predict potential outcomes or treatment effects. A comparison requires the same intervention target, available inputs and identification assumptions. Our joint target may include several outcomes under one intervention. Arm-specific distributions alone do not identify the distribution of an individual's treatment effect: equal Bernoulli marginals permit both $Y(1) = Y(0)$ and $Y(1) = 1 - Y(0)$.

G.4 Scope of the evidence

Our real-data experiments impose recording so that deployment outcomes remain available for evaluation. Natural recording requires a substantive bridge, source support and representative auxiliary records. The clinical, causal and fairness examples explain how those requirements translate to each application. Their role is methodological motivation, and empirical validation remains specific to the experiments reported here.

What the experiments establish. Accurate population weights cannot compensate for arbitrary component error, and estimated weights add fitting and validation variability. TabPFN's Beijing no-shift loss persists with both Original and Native available to the selector (Appendix F.2.2). The unsuccessful Household primary comparison and independent Electricity test bound the positive results (Appendices F.1.2 and F.2.2). Predictive scores, coverage and association measures assess different properties without establishing complete calibration. Joint draws from scalar predictors require a specified coupling. Consistency across different observed subsets additionally requires the compatibility conditions in Appendix B.11.

Intervals condition on fits or vary them according to the replication units defined in the supplement's opening conventions. More predictive draws reduce simulation error, whereas fitting uncertainty requires replication. Conclusions about a selected procedure apply to its tested candidates and validation data. Transfer to new populations, including groups poorly represented in the source, requires separate evidence.