UNIT-LEVEL AGGREGATION BREAKS COVERAGE IN CAUSAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal foundation models (CFMs) show promising results for treatment-effect estimation, with unit-level estimates that are competitive on standard benchmarks. Many causal analyses, however, inform decisions through effects averaged over groups of units, such as the full sample, the treated units, or a subgroup, and for these estimands a CFM is useful only if its credible intervals are calibrated. We examine these intervals and find that their coverage degrades, both in simple synthetic settings and on standard benchmarks. On ACIC 2016, for example, temperature-calibrated 95% intervals from CausalPFN (Balazadeh et al., 2025a) fall below nominal coverage beyond groups of 30 out of 4802 units, and its interval for the full-sample average contains the true value in only 20% of datasets.

We trace this problem to how these models are trained and how their intervals are formed. They output a predictive distribution for one unit at a time and, at inference time, combine the unit-level variances as if the errors on different units were independent. We show formally that posterior uncertainty over the data-generating process makes these errors correlated. As a result, the combined variance is too small by a factor that grows with the size of the group. Empirically, this factor predicts where coverage fails and by how much. We then show that two neural process models, which make joint predictions for a group of units, account for this correlation and substantially improve coverage on synthetic tasks.

1 INTRODUCTION

Causal Foundation Models (CFMs) are pre-trained on simulated causal datasets and estimate treatment effects for a new dataset in a single forward pass. Recent examples, including CausalPFN (Balazadeh et al., 2025a), Do-PFN (Robertson et al., 2025), and CausalFM (Ma et al., 2026), perform competitively on standard benchmarks such as Kang–Schafer (Kang & Schafer, 2007) and the semi-synthetic families of Hill (2011), without fitting a new estimator to each dataset. CFMs typically estimate effects at the unit level, through Conditional Average Treatment Effects (CATEs) that describe the average treatment effect as a function of the observed covariates.

Many causal questions, however, concern effects averaged over a group of units, such as the population average treatment effect (PATE), the average treatment effect on the treated (ATT), and subgroup effects. Clinical trials often focus on subgroup estimands (ICH, 2019; EMA, 2019), drugs are approved for specific subpopulations on the basis of subgroup estimates and their intervals (Taylor et al., 2004), and large policy evaluations report prespecified subgroup effects whose intervals do not rule out meaningful effects (Baicker et al., 2013). We call any average of the CATE over a set of units a *subset-CATE* (see, e.g., Dorie et al., 2022) and define it formally in Section 3. In principle, a CFM can estimate any subset-CATE by averaging its unit-level estimates. One trained model can then target any subgroup without refitting, including the full sample and the treated units, which give analogues of the PATE and ATT.

In applied settings such as medicine, decisions often depend on how uncertain the estimate is, so a CFM is useful for these questions only if its intervals for group averages are calibrated. We show that the intervals current CFMs report for group averages, including the full sample, have no coverage guarantee and fall below nominal on every benchmark we test, and we propose two models that address the problem. Some CFMs report unit-level credible intervals (Balazadeh et al., 2025a; Ma et al., 2026), and CausalPFN's are conservative on our benchmarks. CausalPFN also

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.

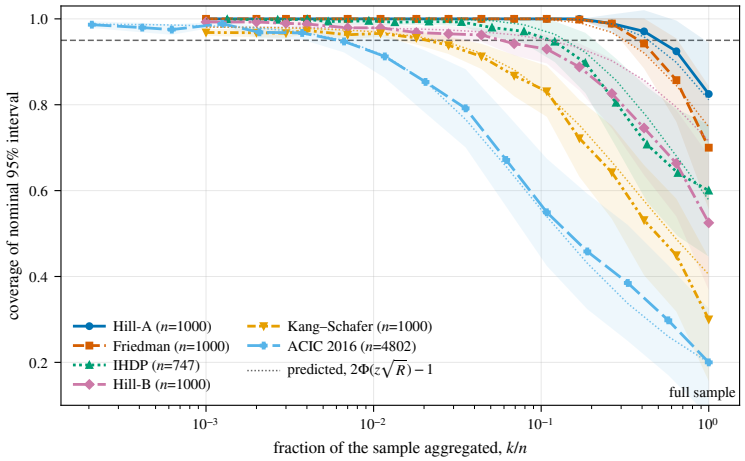

Figure 1: Coverage of nominal 95% subset-CATE intervals from temperature-calibrated CausalPFN on standard benchmarks, with unit-level aggregation. Subsets are drawn uniformly at random. The horizontal axis is the fraction of the sample aggregated, and its right edge is the full sample. Thin dotted lines are the prediction of Equation (1), with m_S and ρ_S measured once per benchmark and held fixed across k, and the horizontal dashed line marks nominal coverage.

returns a full-sample average, which it reports as the ATE, with a 95% credible interval formed by averaging independent draws from the unit-level predictive distributions after an optional temperature calibration. On ACIC 2016 (Dorie et al., 2019), however, this interval contains the true full-sample average in only 8 of the 40 datasets we use when calibrated and in 18 of 40 when not. Mourão et al. (2026) report similarly poor ATE coverage on simulated clinical data. Intervals from the same procedure fall below nominal coverage for subsets larger than 11% of the sample on IHDP (Hill, 2011; Shalit et al., 2017) and 0.6% on ACIC 2016 (Figure 1).

Current CFMs form the aggregate interval by summarizing each unit's uncertainty separately and combining the summaries as if the units' errors were independent; we call this *unit-level aggregation*. Because Do-PFN (Robertson et al., 2025), CausalFM (Ma et al., 2026), and CCPFN (Stith et al., 2026) provide no aggregate estimator, any interval for a subset-CATE has to be built from their unit-level outputs. When there is uncertainty about the data-generating process, however, a model that overestimates one unit's effect tends to overestimate the others' too. Averaging over units then shrinks only the part of the error that is not shared. For a subset of k units, let ρ_S be the correlation between the model's errors on two distinct units of the subset, taken over tasks (draws of the SCM, the dataset, and the subset; Section 4.1), and let m_S be the ratio of the variance the model reports for a unit to its mean squared error ($m_S = 1$ for calibrated unit-level intervals). Under mean-zero Gaussian errors and a fixed reported variance, a nominal 95% interval formed by unit-level aggregation has coverage

$$\mathcal{C}(k) \;=\; 2\,\Phi\!\left(1.96\,\sqrt{\frac{m_S}{1+(k-1)\rho_S}}\right) - 1. \tag{1}$$

The denominator is the design effect of survey sampling (Kish, 1965), applied here to estimation errors. Because it grows linearly in k, coverage tends to zero for any fixed $\rho_S > 0$. With over-wide unit-level intervals ($m_S > 1$), coverage still falls, but at larger k. On ACIC 2016, for example, CausalPFN has $m_S = 1.63$ and $\rho_S = 0.02$. For these values, Equation (1) predicts that coverage falls below nominal at 32 units and reaches 20% on the full sample; we observe 30 units and 20%.

Once the data are observed, the remaining uncertainty in a unit's effect comes from the structural causal model (SCM) that generated them. The effects of all units depend on that same SCM. Unless the data determine the SCM, the unit effects are dependent a posteriori, and an estimator's errors on different units are in general correlated. The exact posterior represents this dependence. Unit-level aggregation ignores it, even though the model's unit-level intervals reflect the same uncertainty.

Accounting for this correlation requires sampling the unit effects jointly. Let τ_i be the CATE of unit i, S the set of units in the subset, ζ the parameters of the SCM, and $\pi(\zeta \mid \mathcal{D})$ their posterior given the

dataset $\mathcal{D}$. Joint and unit-level sampling both average draws of $(\tau_i)_{i\in S}$ and differ only in where the integral over ζ is placed. They draw from

$$\text{unit-level:} \quad \prod_{i\in S} \int p(\tau_i \mid \zeta)\, d\pi(\zeta \mid \mathcal{D}), \tag{2}$$

$$\text{SCM-marginalized:} \quad \int \prod_{i\in S} p(\tau_i \mid \zeta)\, d\pi(\zeta \mid \mathcal{D}). \tag{3}$$

Placing the integral inside the product, as unit-level aggregation does, treats the unit effects as independent once the SCM is integrated out. They share the SCM, however, and this induces the error correlation ρ_S in Equation (1). *SCM-marginalized aggregation* places the integral outside the product and so accounts for the dependence between units. We approximate it with a latent neural process (NP) (Garnelo et al., 2018b), which replaces the SCM with a learned latent variable, and with a direct-decode conditional neural process (CNP) (Garnelo et al., 2018a), which learns the distribution of the subset-CATE directly.

Our contributions are as follows:

- **Diagnosis.** We show that unit-level aggregation understates the variance of any subset-CATE, whether over the full sample, the treated units, or a subgroup, by the design effect. Under Gaussian errors with a fixed reported variance, its coverage is given by Equation (1) (Proposition 1).
- **Source.** We show that, for the posterior mean, the error covariance is the covariance of unit effects under the posterior over the SCM (Proposition 2). It is generally nonzero unless the posterior determines the queried unit effects, and it cannot be computed from unit-level marginals alone.
- **Method.** We propose two SCM-marginalized models, a latent NP and a direct-decode CNP, that integrate over the SCM jointly for all units of the subset (Section 4.3).
- **Evidence.** CausalPFN's intervals lose nominal coverage on all six standard benchmarks, including for the full sample. With the weights held fixed, unit-level coverage at the largest subset is at most 0.47 across 60 cells, while SCM-marginalized coverage has a median of 0.89. Proposition 1 predicts coverage on the synthetic family to within a median absolute error of 0.05 (Section 5).

2 Related work

Inference for aggregate estimands. Uncertainty quantification for group effects, the problem we describe for CFMs in Section 1, has been studied for non-amortized estimators. These can target group effects while retaining the dependence among units, e.g., by averaging within each posterior draw (Hill, 2011), as our latent NP does, or by building the aggregate from orthogonal scores (Chernozhukov et al., 2018a; Semenova & Chernozhukov, 2021; Chernozhukov et al., 2018b). Conformal methods give distribution-free coverage for individual treatment effects (Lei & Candès, 2021; Alaa et al., 2023; Memmesheimer et al., 2025), but target a unit's outcome, not a group average.

Design effects and clustered data. Correlation within groups inflates the variance of an average by the design effect, which Kish (1965) introduced for complex surveys and which also applies to cluster-randomized trials (Hedges, 2007) and domain estimation (Cochran, 1977). We apply it to a model's estimation errors, whose correlation comes from the posterior over the SCM, not from the sampling design.

Uncertainty quantification for PFNs. In the amortized prediction setting, prior-data fitted networks (PFNs) are trained on synthetic learning problems drawn from a prior, so that a single forward pass approximates posterior inference (Müller et al., 2022), and they are state of the art on tabular data (Hollmann et al., 2023; 2025). Nagler (2023) gives a theoretical foundation for PFNs as approximations to the posterior predictive distribution. Later work extends PFNs to full posteriors (Reuter et al., 2025) and to credible sets for a global functional (Ng et al., 2025). Notably, Nagler & Rügamer (2025) name joint inference across query points, which a subset-CATE interval requires, as an open problem. Equations (2) and (3) correspond to the marginal and joint predictive distributions of Wen et al. (2021) and Osband et al. (2023). Gaussian and autoregressive CNPs model the dependence between target points that standard CNPs omit (Markou et al., 2022; Bruinsma et al., 2023).

Causal foundation models. PFNs have been applied to causal effect estimation (Bynum et al., 2025; Balazadeh et al., 2025a; Robertson et al., 2025; Ma et al., 2026; Stith et al., 2026), where they

perform competitively on standard benchmarks. Those that report uncertainty do so for single units or, like CausalPFN, aggregate independent unit-level draws, and two recent papers examine their uncertainty for aggregate effects. Mourão et al. (2026) evaluate CausalPFN on simulated clinical scenarios and report poor coverage of its 95% credible interval for the ATE, attributing it to estimation bias and inadequate uncertainty quantification. Melnychuk et al. (2026) show that a PFN's posterior for an aggregate effect can have a prior-induced bias that does not vanish asymptotically, which precludes frequentist consistency. They also note that a PFN's pointwise predictive distributions do not determine a unique joint posterior and that sampling them independently ignores the dependence between points. They obtain joint draws with martingale posteriors. We quantify the coverage lost by independent aggregation for arbitrary subsets and show that CausalPFN forms its full-sample interval this way. Our latent NP models this dependence through a shared latent variable, and our CNP predicts the subset average directly.

3 SETUP

We consider effect estimation under a binary treatment. Each unit i has covariates $X_i \in \mathcal{X}$, a treatment $T_i \in \{0, 1\}$, and potential outcomes $Y_i(0), Y_i(1)$, one of which is observed as $Y_i = Y_i(T_i)$. We write the CATE as $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$ and $\tau_i := \tau(X_i)$ for its value at unit i.

Subset-CATE. Our target averages these unit-level effects over a queried subset of units. Let $S \subseteq \{1, \dots, n\}$ be the indices of the subset and $k = |S|$ its size. We define

$$\theta_S \;:=\; \frac{1}{k} \sum_{i \in S} \tau_i, \tag{4}$$

and refer to θ_S as the subset-CATE.

Because θ_S averages $\tau(X_i)$ at the realized covariates of the units in S, it is a finite-sample estimand—the estimand of Abadie & Imbens (2006) restricted to the units in S. It differs from the sample average treatment effect $k^{-1} \sum_{i \in S}(Y_i(1) - Y_i(0))$, which includes unit-level noise, and from the PATE $\mathbb{E}[Y(1) - Y(0)]$.

At $k = 1$, θ_S is the CATE of a single unit. At $S = \{1, \dots, n\}$ it is a sample analogue of the PATE, and the quantity existing CFMs report as the ATE; at $S = \{i : T_i = 1\}$ it is the analogous quantity for the ATT.

Meta-learning. CFMs are amortized causal estimators that follow the meta-learning tradition of prior-data fitted networks (Müller et al., 2022; Hollmann et al., 2023). They are trained on tasks $\mathcal{T}_1, \dots, \mathcal{T}_J$ drawn i.i.d. from a task distribution. Each task $\mathcal{T}_j = (\zeta_j, \mathcal{D}_j, S_j)$ consists of a structural causal model (SCM) $\zeta_j \sim \pi_0$, a dataset $\mathcal{D}_j = \{X_i, T_i, Y_i\}_{i=1}^{n_j}$ drawn from the SCM, and a subset S_j of one or more units, drawn given $\mathcal{D}_j$, whose subset-CATE is queried. We write $\tau(x; \zeta)$ for the CATE under SCM ζ when the dependence matters. Current CFMs typically set $|S_j| = 1$, so the training target is the CATE. The model then returns a predictive distribution for a single unit's treatment effect, whose mean and variance we write as $\hat{\tau}_i$ and $\hat{\sigma}_i^2$. Estimating θ_S for $|S| > 1$ with a credible interval requires combining $\hat{\tau}_i$ and $\hat{\sigma}_i^2$ across units in S.

Uncertainty quantification and coverage. CFMs that provide uncertainty quantification typically report a credible interval, since they are designed to approximate a posterior distribution under a known prior. We denote the interval for θ_S as $[\hat{l}_S, \hat{u}_S]$. When it is symmetric about $\hat{\theta}_S$ we write $[\hat{l}_S, \hat{u}_S] = \hat{\theta}_S \pm z_{1-\alpha/2}\, \hat{\sigma}_S$, with $\hat{\sigma}_S$ the standard deviation the model reports for θ_S and $z_{1-\alpha/2}$ the standard normal quantile ($z_{0.975} \approx 1.96$ at our $\alpha = 0.05$).

We evaluate credible intervals by their coverage, the fraction of queries where the interval contains the true subset-CATE. Coverage at subset size k is

$$\mathcal{C}(k) \;:=\; \Pr\big(\theta_S \in [\hat{l}_S, \hat{u}_S]\big), \tag{5}$$

where the probability is over the SCM $\zeta \sim \pi$, the dataset $\mathcal{D} \mid \zeta$, and the subset S of size k, drawn uniformly or by a selection rule. Here π is the distribution of evaluation tasks, which can differ from a model's training prior π_0; for a benchmark, the probability is over its datasets. Appendix B gives the estimator we use for $\mathcal{C}(k)$ and the task counts behind each reported value.

Identification. Our analysis assumes that all confounders are observed and that positivity holds, so that the subset-CATE is identified. Our synthetic family satisfies both by construction (Appendix A). The benchmarks have no unobserved confounding, but IHDP and some ACIC 2016 settings have limited overlap, so part of their error can come from extrapolation.

4 WHY UNIT-LEVEL AGGREGATION BREAKS COVERAGE

4.1 THE DESIGN EFFECT

Let $e_i := \hat{\tau}_i - \tau_i$ denote the estimator's error at unit i. Since θ_S is an average of unit-level effects, the subset-CATE error is the average of unit-level errors, $\bar{e}_S = \hat{\theta}_S - \theta_S = k^{-1} \sum_{i \in S} e_i$. We define two moments of the errors and their ratio, with expectations over tasks $(\zeta, \mathcal{D}, S)$ as in Equation (5),

$$\sigma_S^2 \;=\; \mathbb{E}\Big[\frac{1}{k} \sum_{i \in S} e_i^2\Big], \qquad c_S \;=\; \mathbb{E}\Big[\frac{1}{k(k-1)} \sum_{\substack{i,j \in S \\ i \neq j}} e_i e_j\Big], \qquad \rho_S \;=\; \frac{c_S}{\sigma_S^2}. \tag{6}$$

We use raw second moments because coverage depends on the total error, including any bias in unit-level estimates. We define the moments over the queried subset because a query that selects units on a covariate (e.g., patients aged 40–60) may have more strongly correlated errors than the dataset as a whole. When S is drawn uniformly, σ_S^2, c_S, and m_S equal their values for the full sample and so do not depend on k. Appendix D shows how ρ_S and m_S change for selected subgroups.

Existing CFMs combine the unit-level variances as if the errors were independent, which gives the reported variance

$$\hat{\sigma}_S^2 \;=\; \frac{1}{k^2} \sum_{i \in S} \hat{\sigma}_i^2. \tag{7}$$

CausalPFN's aggregate estimator averages independent draws from the units' predictive distributions, which makes it a Monte Carlo version of Equation (7) (Appendix E).[1]

Proposition 1 (Design effect; cf. Kish 1965)**.** *Let $\hat{\theta}_S = k^{-1} \sum_{i \in S} \hat{\tau}_i$, with expectations over ζ, $\mathcal{D}$, and S as in Equation* (5)*. The mean squared error of $\hat{\theta}_S$ is*

$$\mathbb{E}\big[\bar{e}_S^2\big] \;=\; \frac{1}{k^2}\, \mathbb{E}\Big[\, \sum_{i \in S} e_i^2 \;+\; \sum_{\substack{i,j \in S \\ i \neq j}} e_i e_j \,\Big] \;=\; \frac{\sigma_S^2}{k}\Big[\, 1 + (k-1)\rho_S \,\Big]. \tag{8}$$

If the interval is $\hat{\theta}_S \pm z_{1-\alpha/2} \hat{\sigma}_S$ with $\hat{\sigma}_S^2$ from Equation (7)*, then the ratio of reported to actual squared error and (when $\hat{\sigma}_S$ does not vary across queries) the coverage are*

$$R \;:=\; \frac{\mathbb{E}[\hat{\sigma}_S^2]}{\mathbb{E}[\bar{e}_S^2]} \;=\; \frac{m_S}{1 + (k-1)\rho_S}, \tag{9}$$

$$\mathcal{C}(k) \;=\; \Pr\left(\frac{|\bar{e}_S|}{\sqrt{\mathbb{E}[\bar{e}_S^2]}} \;\leq\; z_{1-\alpha/2} \sqrt{R} \right), \tag{10}$$

where $m_S := \mathbb{E}[\bar{\hat{\sigma}}_S^2]/\sigma_S^2$ and $\bar{\hat{\sigma}}_S^2 = k^{-1} \sum_{i \in S} \hat{\sigma}_i^2$ is the average variance the model reports for a unit in S. If, in addition, $\bar{e}_S$ is mean-zero Gaussian, then $\mathcal{C}(k) = 2\Phi(z_{1-\alpha/2}\sqrt{R}) - 1$.

Proof sketch (Appendix G). Expanding $\bar{e}_S^2$ gives k diagonal terms, which contribute σ_S^2 on average, and $k(k-1)$ cross terms, which contribute c_S; the rest follows by dividing Equation (7) by Equation (8) and standardizing the error.

The standardized error in Equation (10) has second moment one regardless of k, ρ_S, and m_S. As a result, once the shape of its distribution is fixed, these quantities affect coverage only through R. Under Gaussian errors, the interval attains nominal coverage at $R = 1$, and for $R < 1$ coverage is

[1] `src/causalpfn/causal_estimator.py:517` and `src/causalpfn/models/icl_model.py:61` of Balazadeh et al. (2025b), at commit `7da4afa`.

below nominal and increasing in R. For $\rho_S > 0$, setting $R = 1$ gives the largest subset for which the reported interval is wide enough,

$$k^* \;=\; 1 + \frac{m_S - 1}{\rho_S}. \tag{11}$$

Calibrated unit-level intervals ($m_S = 1$) give $k^* = 1$, so nominal coverage holds for single units only. Narrower ones ($m_S < 1$) attain nominal coverage at no subset size under Gaussian errors. Wider ones attain it up to k^*, with $k^* - 1$ inversely proportional to ρ_S.

Equations (8) and (9) hold for any error distribution. When the reported variance varies across queries, Equation (1) is an approximation, since R is a ratio of expectations (Appendix G); we test it in Section 5.2.

4.2 The source of the error correlation

For the posterior mean, the correlation ρ_S in Proposition 1 comes entirely from posterior uncertainty over the SCM. Write $Z_i = (X_i, T_i, Y_i)$ for the observed record of unit i. Given the SCM, the units of a task are i.i.d. The SCM itself is random, however, so the joint distribution of the units is a mixture over SCMs,

$$p(Z_1, \ldots, Z_n) = \int \prod_{i=1}^{n} p(Z_i \mid \zeta)\, d\pi(\zeta), \tag{12}$$

which does not factorize across units. Because observing one unit is informative about ζ, and hence about the others, the units are exchangeable but not independent. De Finetti's theorem (de Finetti, 1937), in the general form of Hewitt & Savage (1955), shows that any infinitely exchangeable sequence admits such a representation for some random measure. For CFMs, that measure is the observational distribution of the SCM, and under the assumptions of Section 3 each $\tau(X_i; \zeta)$ is a functional of it. A CFM that is permutation invariant in its context treats its data as exchangeable.

Proposition 2 (Source of the correlation)**.** *For any estimator whose outputs are functions of $\mathcal{D}$,*

$$\mathrm{Cov}\left(e_i, e_j \mid \mathcal{D}\right) \;=\; \mathrm{Cov}_{\zeta \sim \pi(\zeta \mid \mathcal{D})}\left(\tau(X_i; \zeta),\, \tau(X_j; \zeta)\right). \tag{13}$$

If S is conditionally independent of the unit effects given $\mathcal{D}$, then c_S is the average, over $\mathcal{D}$, S, and pairs of distinct units in S, of this covariance plus $b_i b_j$, where $b_i := \mathbb{E}[e_i \mid \mathcal{D}]$. For the posterior mean under π, $b_i = 0$.

Proof sketch (Appendix G). Given $\mathcal{D}$, the outputs $\hat{\tau}_i$ are fixed; given ζ as well, each τ_i is fixed. Hence $\mathrm{Cov}(e_i, e_j \mid \zeta, \mathcal{D}) = 0$, and the law of total covariance leaves only the covariance of τ_i and τ_j across posterior draws of ζ. Averaging $\mathbb{E}[e_i e_j \mid \mathcal{D}] = \mathrm{Cov}(e_i, e_j \mid \mathcal{D}) + b_i b_j$ over $\mathcal{D}$, S, and pairs gives c_S. The term $b_i b_j$ is positive when the estimator errs in the same direction at units i and j.

Equation (3) includes this covariance, and Equation (2) sets it to zero. The unit-level marginals $p(\tau_i \mid \mathcal{D})$ do not determine it (Appendix G), so recovering it requires a joint model of the unit effects or, as in martingale posteriors (Ng et al., 2025; Melnychuk et al., 2026), repeated forward passes.

4.3 SCM-marginalized aggregation

The posterior variance of θ_S averages the covariance of Proposition 2 over all k^2 pairs of units,

$$\mathrm{Var}\left(\theta_S \mid \mathcal{D}\right) \;=\; \frac{1}{k^2} \sum_{i \in S} \sum_{j \in S} \mathrm{Cov}_{\zeta \sim \pi(\zeta \mid \mathcal{D})}\left(\tau(X_i; \zeta),\, \tau(X_j; \zeta)\right). \tag{14}$$

In practice, both distributions in Equations (2) and (3) are approximated by Monte Carlo. The two schemes differ in whether the unit effects in one draw are sampled independently or computed under one shared draw of the SCM. Let $\tilde{\theta}_S = k^{-1} \sum_{i \in S} \tilde{\tau}_i$ be one draw of the subset-CATE and $\tilde{\zeta} \sim \pi(\zeta \mid \mathcal{D})$ one draw of the SCM. Conditional on $\mathcal{D}$, the variance of one draw is

$$\text{unit-level:} \quad \tilde{\tau}_i \overset{\text{ind.}}{\sim} p(\tau_i \mid \mathcal{D}), \qquad \mathrm{Var}\left(\tilde{\theta}_S \mid \mathcal{D}\right) = \tfrac{1}{k^2} \textstyle\sum_{i \in S} \mathrm{Var}(\tau_i \mid \mathcal{D}), \tag{15}$$

$$\text{SCM-marginalized:} \quad \tilde{\tau}_i = \tau(X_i; \tilde{\zeta}), \qquad \mathrm{Var}\left(\tilde{\theta}_S \mid \mathcal{D}\right) = \mathrm{Var}(\theta_S \mid \mathcal{D}). \tag{16}$$

dataset	covariates	source	n	d
Synthetic family	simulated $\mathcal{N}(0, I)$	Appendix A	500	3
Hill-A	simulated $\mathcal{N}(0, I)$	Hill (2011), setting A	1000	25
Friedman	simulated $\mathrm{U}(0, 1)$	Friedman (1991), confounded	1000	10
IHDP	real, IHDP study	Hill (2011); splits of Shalit et al. (2017)	747	25
Hill-B	simulated $\mathcal{N}(0, I)$	Hill (2011), setting B	1000	25
Kang–Schafer	simulated, nonlinear in 4 latents	Kang & Schafer (2007)	1000	4
ACIC 2016	real, competition covariate table	Dorie et al. (2019)	4802	58

Table 1: Datasets used in our experiments. Our models are trained and evaluated on the synthetic family, and CausalPFN is evaluated on all seven. Hill-A, Hill-B, Friedman, and Kang–Schafer are our reimplementations (Appendix A). d is the number of covariates supplied to the estimator.

The first is what CausalPFN's aggregate estimator computes, and it omits the cross terms of Equation (14). The second includes them, because the units in a draw share one SCM. For the exact posterior under π, the first gives $R = 1/\big(1 + (k-1)\rho_S\big)$ in Proposition 1, and the second gives $R = 1$ and exact coverage at every k (Corollary 1). Since no CFM outputs draws of ζ, the models below approximate SCM-marginalized aggregation without them.

Latent NP. A latent NP (Garnelo et al., 2018b) replaces the SCM with a learned latent variable z. An encoder maps the context to a distribution $q_\psi(z \mid \mathcal{D})$, and a decoder maps z, the context, and a unit's covariates to a distribution over that unit's effect. The model's joint distribution over the unit effects is then

$$p_\phi\big((\tau_i)_{i\in S} \mid \mathcal{D}\big) \;=\; \int \prod_{i\in S} p_\phi(\tau_i \mid z, X_i, \mathcal{D})\, q_\psi(z \mid \mathcal{D})\, dz, \tag{17}$$

which has the SCM-marginalized form of Equation (3) with z in place of ζ. The model is trained on two likelihood terms estimated with draws of z from $q_\psi(z \mid \mathcal{D})$, with no KL term (cf. Foong et al., 2020): one for each unit's effect and one for θ_S on a random subset S, averaging the decoded unit effects within each draw (Appendix B). Because the decoder treats units as independent given z, any dependence between units that the model learns must be captured by z. The subset-level term trains this dependence directly, since for large subsets the variance the model assigns to θ_S comes mainly from how the within-draw averages vary across draws. At inference, we sample one z per Monte Carlo draw, decode every unit of S from it, and average, which implements Equation (16) with z in place of ζ. Uncertainty in the per-unit decoder given z is still combined as in unit-level aggregation.

Direct-decode CNP. The CNP (Garnelo et al., 2018a) takes S as a mask in the forward pass and outputs a Gaussian $q_\phi(\theta_S \mid \mathcal{D}, S)$. When it is trained to minimize $\mathbb{E}[-\log q_\phi(\theta_S \mid \mathcal{D}, S)]$ over uniformly drawn subsets of all sizes, the best Gaussian output has the mean and variance of the posterior $\int p(\theta_S \mid \zeta)\, d\pi_0(\zeta \mid \mathcal{D})$ under its training prior π_0 (Proposition 3). It thus reports the variance of Equation (14), cross terms included, without sampling. Because S is an input, the model need not be retrained for new subsets.

5 Experiments

5.1 Setup

Datasets. We train and evaluate our own models on a synthetic family of SCMs (Appendix A). Each task draws a new SCM with three covariates and $n = 500$ units, and θ_S has a closed form for every subset. A sharpness parameter s controls how far propensity scores are pushed toward 0 and 1; we use $s = 1$ and $s = 4$ for our models and $s \in \{1, 2, 4, 6\}$ for CausalPFN. CausalPFN is evaluated on the same family and on six standard benchmarks (Table 1). Because all of them have simulated response surfaces, we score intervals against each unit's known conditional effect (Appendix A).

Models. CausalPFN is used as released, without retraining, with intervals formed by its own aggregate estimator. On the benchmarks we enable its built-in per-dataset temperature calibration; on the synthetic family we use its default, which leaves calibration off. Our CNP and latent NP are trained at $s = 1$, with three seeds each (Appendix B).

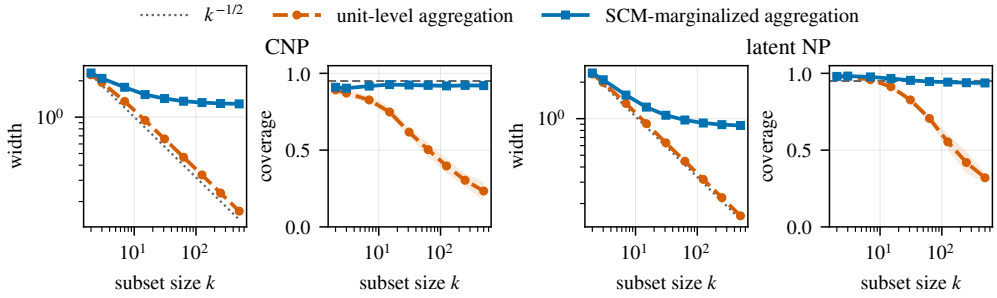

Figure 2: Unit-level and SCM-marginalized aggregation from one CNP and one latent NP, with uniformly drawn subsets at $s = 1$, showing interval width and coverage against subset size.

Experiments. The *benchmark* experiment applies CausalPFN to each benchmark, querying uniformly drawn subsets of increasing size on 40 datasets per benchmark. The *fixed-weight* experiment holds a trained CNP or latent NP fixed and, for the same subset, forms one interval by unit-level aggregation, as in Equation (2), and one by SCM-marginalized aggregation, as in Equation (3). Because weights, context, subsets, and targets are shared, the comparison isolates the aggregation step (for the CNP, the point estimate also differs). It tests the mechanism and is not a comparison with CausalPFN. It covers six subset selection rules for the CNP and three for the latent NP, plus $s = 4$ for the CNP with uniformly drawn subsets. We call each of the 60 combinations of trained model, rule, subgroup size, and s a *cell* (Appendix B). The *variance-ratio* experiment compares coverage with Proposition 1 for every model, rule, and subset size.

Metrics. We report coverage of nominal 95% intervals, estimated over tasks and subsets as in Appendix B, and $\hat{k}^*$, the subset size at which observed coverage first falls below nominal. Because m_S and ρ_S require the true effects, we use them only to explain the results.

5.2 Results

Benchmarks. Table 4 in Appendix F gives the full results behind Figure 1. Coverage starts at or above nominal on every benchmark in the figure and falls below it as subsets grow, fastest on ACIC 2016 and slowest on Hill-A, where it falls below nominal only beyond half the sample. Equation (1) approximates the observed curves but overstates coverage on Hill-B and Kang–Schafer (Appendix F). The error correlation alone does not predict where coverage fails. Hill-A has the largest correlation in the suite ($\rho_S = 0.077$) but is the last to fall below nominal, while Hill-B, with a correlation twenty times smaller, loses coverage at 5.8% of its sample. Equation (11) accounts for the ordering, since Hill-A's unit-level intervals are far too wide ($m_S = 35$) and Hill-B's only slightly ($m_S = 1.42$). Accurate unit-level estimates do not ensure aggregate coverage either. Although Kang–Schafer has the second-lowest normalized RMSE in the suite (0.09), it loses coverage at 2.1% of its sample.

Fixed weights. In Figure 2, the unit-level interval shrinks at close to the $k^{-1/2}$ rate that independence implies, and at the largest subsets it is 7.7 times narrower than the SCM-marginalized interval for the CNP and 5.5 times for the latent NP. Unit-level coverage falls with k, while SCM-marginalized coverage stays between 0.90 and 0.98. Across all 60 cells, SCM-marginalized coverage at the largest subset has a median of 0.89 (range 0.70 to 0.99), and its lowest value exceeds the highest unit-level value (0.47). Coverage is lowest for the CNP at $s = 4$, outside its training distribution, and for subsets selected on the true effect, which Corollary 1 excludes because S depends on the SCM. The latent NP, whose decoder treats units as independent given z, has a median of 0.83. In a separate comparison without fixed weights, with three models trained at $n = 2000$, unit-level coverage is 0.17 at $k = n$, and the two SCM-marginalized models reach 0.97 and 0.98 (Table 5).

Variance ratio. Figure 3 plots observed coverage against R. We compare it with the prediction for $R < 1$, since for $R \geq 1$ the predicted coverage is at least nominal. Unit-level coverage closely follows Equation (1), with a median absolute error of 0.043 for our models and 0.021 for CausalPFN on the synthetic family. Because Equations (8) and (9) are exact given m_S and ρ_S, the figure mainly

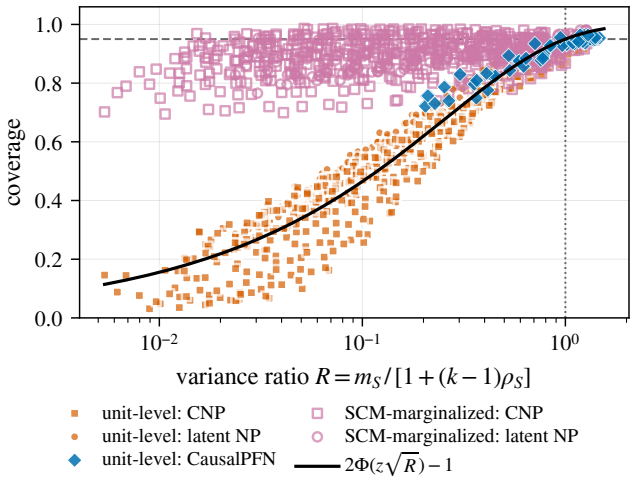

Figure 3: Measured coverage against the variance ratio R of Proposition 1, pooled across models, values of s, subset selection rules, and subset sizes. The curve is Equation (1), with m_S and ρ_S measured from unit-level errors. SCM-marginalized points are plotted at the R of unit-level aggregation.

tests the Gaussian and fixed-variance assumptions. SCM-marginalized points lie well above the curve (median absolute error 0.315), since their intervals include the cross terms.

6 Discussion and Limitations

Our CNP and latent NP are small models trained on a synthetic family; we do not claim they match CausalPFN's accuracy or generality. We evaluate them only on that family, since every benchmark has more covariates and units. Because our family has few parameters, a small latent can capture most of the posterior dependence, but richer priors may require a larger one. On benchmarks, part or all of ρ_S can come from the shared bias term $b_i b_j$ of Proposition 2, which no aggregation rule removes.

For existing models, rescaling unit-level intervals to be calibrated did not help in our experiments, and it made coverage worse when they were too wide (Appendix C). Multiplying the aggregate variance by the design effect brings coverage close to nominal, but it requires ρ_S and m_S. They could be estimated on simulated tasks, but they vary across our benchmarks by more than an order of magnitude (Table 4) and depend on the model as well as the subgroup.

7 Conclusion

Causal foundation models form intervals for effects averaged over the full sample, the treated units, or a subgroup by combining unit-level intervals as if the errors were independent. Because posterior uncertainty over the SCM makes these errors correlated, the combined variance is too small by the design effect, and coverage falls as the group grows. We observe this on all six standard benchmarks, trace it to the aggregation step with the weights held fixed, and predict its size from m_S and ρ_S. Integrating over the SCM jointly for all units recovers the missing covariance. Both models we test give much higher coverage than unit-level aggregation, though not always at the nominal level.

Use of AI tools

We used large language models to aid and polish writing, for retrieval and discovery of related work, and to draft sections of this paper. All drafted text was reviewed and revised by the authors, all cited work was verified against primary sources, and all theoretical claims and experimental results are the authors' own.

Reproducibility statement

Appendix A specifies the synthetic family in full, including the coefficient priors and the propensity mechanism, and gives the covariate source, sample size, and generating design for each of the six benchmarks. Appendix B gives the architectures, the training setup and checkpoint selection for each model, and the task and subset counts behind every coverage estimate we report. Appendix E identifies the CausalPFN code paths our characterization relies on, at a pinned commit. CausalPFN is used as released, without retraining, with the calibration settings of Section 5.1. Code to reproduce every figure and table, with the recorded configuration and training command for each of our six trained models, is included in the supplementary material.

References

Alberto Abadie and Guido W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Ahmed M. Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. In *Advances in Neural Information Processing Systems*, pp. 47682–47703, 2023. Oral. arXiv:2308.14895.

Katherine Baicker, Sarah L. Taubman, Heidi L. Allen, Mira Bernstein, Jonathan H. Gruber, Joseph P. Newhouse, Eric C. Schneider, Bill J. Wright, Alan M. Zaslavsky, and Amy N. Finkelstein. The Oregon experiment: Effects of Medicaid on clinical outcomes. *New England Journal of Medicine*, 368(18):1713–1722, 2013.

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*, 2025a. Spotlight. arXiv:2506.07918.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN, 2025b. Software repository, commit `7da4afa`. `https://github.com/vdblm/CausalPFN`.

Wessel P. Bruinsma, Stratis Markou, James Requeima, Andrew Y. K. Foong, Tom R. Andersson, Anna Vaughan, Anthony Buonomo, J. Scott Hosking, and Richard E. Turner. Autoregressive conditional neural processes. In *International Conference on Learning Representations*, 2023. arXiv:2303.14468.

Lucius EJ Bynum, Aahlad Manas Puli, Diego Herrero-Quevedo, Nhi Nguyen, Carlos Fernandez-Granda, Kyunghyun Cho, and Rajesh Ranganath. Black box causal inference: Effect estimation via meta prediction. *arXiv preprint arXiv:2503.05985*, 2025.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments. *arXiv preprint arXiv:1712.04802*, 2018b.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3rd edition, 1977.

Bruno de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68, 1937.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Vincent Dorie, George Perrett, Jennifer L. Hill, and Benjamin Goodrich. Stan and bart for causal inference: Estimating heterogeneous treatment effects using the power of stan and the flexibility of machine learning. *Entropy*, 24(12), 2022. ISSN 1099-4300. doi: 10.3390/e24121782. URL `https://www.mdpi.com/1099-4300/24/12/1782`.

EMA. Guideline on the investigation of subgroups in confirmatory clinical trials. Technical Report EMA/CHMP/539146/2013, European Medicines Agency, 2019.

Andrew Y. K. Foong, Wessel P. Bruinsma, Jonathan Gordon, Yann Dubois, James Requeima, and Richard E. Turner. Meta-learning stationary stochastic process prediction with convolutional neural processes. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2007.01332.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Marta Garnelo, Dan Rosenbaum, Christopher J. Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J. Rezende, and S. M. Ali Eslami. Conditional neural processes. In *International Conference on Machine Learning*, volume 80 of *PMLR*, pp. 1704–1713, 2018a. arXiv:1807.01613.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b. ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models.

Larry V. Hedges. Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32(4):341–370, 2007.

Edwin Hewitt and Leonard J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023. arXiv:2207.01848.

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

ICH. ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials. Technical Report EMA/CHMP/ICH/436221/2017, International Council for Harmonisation, 2019.

Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B*, 83(5):911–938, 2021. arXiv:2006.06138.

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*, 2026. arXiv:2506.10914.

Stratis Markou, James Requeima, Wessel P. Bruinsma, Anna Vaughan, and Richard E. Turner. Practical conditional neural processes via tractable dependent predictions. In *International Conference on Learning Representations*, 2022. arXiv:2203.08775.

Valentyn Melnychuk, Vahid Balazadeh, Stefan Feuerriegel, and Rahul G. Krishnan. Frequentist consistency of prior-data fitted networks for causal inference. In *International Conference on Machine Learning*, 2026. Poster. arXiv:2603.12037.

Marcel Memmesheimer, Cyrill Heuveline, and Jürgen Hesser. A systematic review of conformal inference procedures for treatment effect estimation: Methods and challenges. *arXiv preprint arXiv:2509.21660*, 2025.

Francisco Mourão, David Hajage, Daria Bystrova, Bertrand Bouvarel, Nathanaël Lapidus, Fabrice Carrat, and Benjamin Glemain. Prior-data fitted networks for causal inference: a simulation study with real-world scenarios. *arXiv preprint arXiv:2603.15928*, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations*, 2022. arXiv:2112.10510.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 25660–25676, 2023. arXiv:2305.11097.

Thomas Nagler and David Rügamer. Uncertainty quantification for prior-data fitted networks using martingale posteriors, 2025. arXiv:2505.11325.

Kenyon Ng, Edwin Fong, David T. Frazier, Jeremias Knoblauch, and Susan Wei. TabMGP: Martingale posterior with TabPFN. *arXiv preprint arXiv:2510.25154*, 2025.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2107.08924.

Arik Reuter, Tim G. J. Rudner, Vincent Fortuin, and David Rügamer. Can transformers learn full Bayesian inference in context? In *International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 51531–51582, 2025. arXiv:2501.16825.

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. Robertson and Reuter contributed equally. arXiv:2506.06039.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021. arXiv:1702.06240.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 3076–3085, 2017. arXiv:1606.03976.

Christopher Stith, Medha Barath, Vahid Balazadeh, Jesse C. Cresswell, and Rahul G. Krishnan. Causal foundation models with continuous treatments, 2026. arXiv:2605.15133.

Anne L. Taylor, Susan Ziesche, Clyde Yancy, Peter Carson, Ralph D'Agostino, Keith Ferdinand, Malcolm Taylor, Kirkwood Adams, Michael Sabolinski, Manuel Worcel, and Jay N. Cohn. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *New England Journal of Medicine*, 351(20):2049–2057, 2004.

Zheng Wen, Ian Osband, Chao Qin, Xiuyuan Lu, Morteza Ibrahimi, Vikranth Dwaracherla, Mohammad Asghari, and Benjamin Van Roy. From predictions to decisions: The importance of joint predictive distributions. *arXiv preprint arXiv:2107.09224*, 2021.

A Data-generating processes

We use a synthetic family so that the true subset-CATE can be computed exactly for any subset and the conditions under which coverage fails can be varied.

A.1 Synthetic family

Each task draws a fresh set of coefficients, then n units from the resulting SCM. Covariates are $X_\ell \sim \mathcal{N}(0,1)$ independently for $\ell = 1, \dots, d$ with $d = 3$, and each unit is summarized by the feature vector

$$h(X) = [\, X_1, \dots, X_d,\ X_1^2, \dots, X_d^2 \,] \in \mathbb{R}^6, \tag{18}$$

which includes squared terms so that both treatment assignment and outcome may depend nonlinearly on the covariates.

Treatment is assigned according to a propensity score $p(X) = \Pr(T = 1 \mid X)$, and the outcome is generated as

$$T \mid X \sim \text{Bernoulli}\big(p(X)\big), \qquad p(X) = 0.05 + 0.90\,\text{expit}\big(s\,h(X)^\top w_p\big), \tag{19}$$

$$Y = h(X)^\top w_b + T\,\tau(X) + u, \qquad u \sim \mathcal{N}(0, 0.25^2). \tag{20}$$

The sharpness s controls how far propensity scores are pushed toward 0 and 1. For any s, $p(X)$ stays in $[0.05, 0.95]$, so positivity holds. Our models are trained only at $s = 1$, and the CNP is also evaluated at $s = 4$, outside its training distribution. We evaluate CausalPFN at $s \in \{1, 2, 4, 6\}$, and the settings above $s = 1$ account for 25 of its 29 points with $R < 1$ in Figure 3.

The CATE is

$$\tau(X) = 0.25 + h(X)^\top w_e. \tag{21}$$

The covariates enter both the propensity score and the outcome, so they are confounders. Since they are fully observed, τ is identified.

Since $\tau(X)$ is available in closed form, θ_S is computed exactly for every task and subset, and targets add no Monte Carlo error to the coverage estimates.

The per-task coefficients are drawn as

$$w_b \sim \mathcal{N}(0, 0.5^2 I_6), \qquad w_e \sim \mathcal{N}(0, 0.35^2 I_6), \qquad w_p \sim \mathcal{N}(0, 0.35^2 I_6), \tag{22}$$

independently across tasks.

Given the coefficients, units are drawn independently, so they are i.i.d. given the SCM ζ and exchangeable but not independent marginally, as in Section 4.2.

A.2 Benchmarks

We evaluate CausalPFN on the six standard benchmarks of Table 1 and on a heavy-tailed variant of Hill-B. In each, the response surfaces are simulated, so each unit's conditional effect $\mu_1(X_i) - \mu_0(X_i)$ is known. We score intervals against it, not against the difference of the two noisy potential outcomes, and there is no unobserved confounding by construction. IHDP and ACIC 2016 pair the simulated surfaces with covariates from real studies; IHDP uses the `ihdp_npci_1-100` splits and ACIC 2016 the official `aciccomp2016` generator. The other designs simulate the covariates as well, with $n = 1000$ and independent $\mathcal{N}(0, 1)$ noise on each potential outcome.

- **Hill-A.** Twenty-five $\mathcal{N}(0, 1)$ covariates, a logistic propensity in the first five, $\mu_0 = X\beta$, and $\mu_1 = X\beta + 4$ with $\beta \sim \mathcal{N}(0, 0.5^2 I)$, so the effect is constant at 4, as in surface A of Hill (2011).
- **Hill-B.** The same covariates and propensity, $\mu_0 = \exp\big((X + 0.5)\beta_B\big)$ with the entries of β_B drawn from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, and $\mu_1 = X\gamma + c$ with $\gamma \sim \mathcal{N}(0, 0.3^2 I)$ and c set so that the average effect on the treated is 4. Unlike Hill (2011), γ is drawn separately from β_B. The heavy-tailed variant replaces the Gaussian noise with t_3 noise.
- **Friedman.** Ten $\mathrm{U}(0, 1)$ covariates, a logistic propensity in $2(X_1 - 0.5) + 1.5(X_3 - 0.5)$, $\mu_0 = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5$ as in Friedman (1991), and effect $2 + X_1 - X_2$.
- **Kang–Schafer.** Four latent variables $Z \sim \mathcal{N}(0, I)$, of which the estimator sees only the transformed covariates of Kang & Schafer (2007), a logistic propensity in Z, and $\mu_0 = \mu_1 = 210 + 27.4Z_1 + 13.7(Z_2 + Z_3 + Z_4)$, so the effect is zero for every unit.

We use splits 1–40 of 100 for IHDP, settings 1–40 of 77 for ACIC 2016 (one replication each), and seeds 1–40 for the simulated designs, for 40 datasets per benchmark. In Hill-A, Friedman, and Kang–Schafer the effect is the same function of the covariates in every dataset, so under the benchmark's own distribution the posterior covariance in Proposition 2 is zero and all of ρ_S comes from the shared bias term $b_i b_j$.

B TRAINING AND EVALUATION PROTOCOL

Architectures. Both models encode each unit's covariates, treatment, and outcome, and mean-pool the encodings into a summary of the sample. The CNP also pools the encodings of the queried subset and decodes a Gaussian for θ_S from the two summaries in a single pass. The latent NP adds a 16-dimensional Gaussian latent z whose mean and scale are computed from the sample summary, decodes each unit's effect given z as a 64-bin histogram on $[-4, 4]$ in standardized units, and forms θ_S within each draw of z. Every network is a two-layer ReLU MLP of width 128, and neither model uses attention.

Training. Both models are trained on the synthetic family of Appendix A at $s = 1$ and $n = 500$. The CNP is trained on subsets drawn uniformly at random, with size sampled log-uniformly between 1 and n, by Gaussian negative log-likelihood. The latent NP's loss uses 16 draws of z and has two terms. The first is the cross-entropy of each unit's histogram against a Gaussian-smoothed target at its true effect, averaged over draws. The second, with weight 2, is the negative log of the average over draws of a Gaussian likelihood for θ_S, with S drawn as for the CNP and, within each draw, the Gaussian matching the mean and variance of the average of the decoded unit effects. There is no KL term, and z is drawn from the same distribution at inference, so this is a Monte Carlo likelihood objective similar to that of Foong et al. (2020). The latent NP standardizes each task's outcome and covariates, and the CNP works on the raw scale, in both training and evaluation. We use Adam with learning rate 3×10^{-4} and cosine decay to 3×10^{-6}, batches of 64 tasks, and 30,000 steps. We train three seeds of each model. One seed takes about one GPU-hour for the CNP and under two for the latent NP.

Subset selection rules. The fixed-weight experiment uses six rules, each applied to a sample of $n = 500$: uniformly drawn subsets of the whole sample; the top 25% of units by X_1; the bottom fraction by true propensity score; a random anchor unit and its nearest neighbours in covariate space; the top fraction by the model's own predicted effect; and the top fraction by the true effect, an oracle rule. For the last four rules, the fraction is 0.10, 0.25, or 0.50. The CNP uses all six rules at $s = 1$ and uniformly drawn subsets at $s = 4$, for 15 settings; the latent NP uses the first three rules at $s = 1$, for 5 settings. With three seeds each, this gives 45 CNP cells and 15 latent NP cells. Uniformly drawn subsets vary k at a fixed error correlation, and the other rules select subgroups, as an analyst would. For the latent NP, the two intervals share the unit-level marginals and differ only in the aggregation step; for the CNP, they also differ in the point estimate, since the SCM-marginalized interval is decoded directly.

Checkpoint selection. We evaluate the checkpoint with the best score over the full run on 256 held-out validation tasks, using negative log-likelihood for the CNP and per-unit RMSE for the latent NP.

Estimating coverage. We estimate $\mathcal{C}(k)$ of Equation (5) over M tasks with L subsets sampled per task,

$$\widehat{\mathcal{C}}(k) \;=\; \frac{1}{ML} \sum_{j=1}^{M} \sum_{l=1}^{L} \mathbf{1}\big\{\theta_{S_{jl}} \in [\hat{l}_{S_{jl}}, \hat{u}_{S_{jl}}]\big\}. \tag{23}$$

Standard errors on $\widehat{\mathcal{C}}(k)$ are clustered by task, since subsets from one task share its error component. The fixed-weight experiment uses 64 tasks with 24 subsets each, the $n = 2000$ comparison 96 tasks with 24 subsets each, and the benchmarks 40 datasets with 48 subsets each.

Measuring ρ_S and m_S. We compute σ_S^2 and c_S as the empirical second moments of Equation (6), from unit-level errors against the analytic targets of Appendix A, and m_S as the model's average reported unit-level variance divided by σ_S^2. Both ρ_S and m_S require the true effects, so neither can be computed in practice, and we use them only as diagnostics.

model	data	m_S	$k_{\max}$	interval	cov, smallest k	cov, $k_{\max}$
CausalPFN	$s = 1$, all units	21.90	500	reported	1.00	0.94
				rescaled	0.91	0.41
				+ design effect	0.91	0.93
CausalPFN	$s = 4$, bottom 10%	3.63	44	reported	0.99	0.79
				rescaled	0.96	0.58
				+ design effect	0.96	0.93
CausalPFN	IHDP	5.43	747	reported	1.00	0.60
				rescaled	0.98	0.25
				+ design effect	0.98	0.93
CNP (ours)	$s = 1$	0.93	500	unit-level	0.89	0.23
				rescaled	0.91	0.24
				+ design effect	0.92	0.93

Table 2: Coverage of the three intervals of Appendix C at the smallest and largest subset in each setting. $k_{\max}$ is the largest evaluated subset size within the pool. The + *design effect* interval uses ρ_S and m_S measured against the true effects and is not available in practice.

C Correcting for the design effect

Correcting m_S and then ρ_S separates the contribution of unit-level calibration from that of the aggregation step.

Table 2 compares three intervals built from the same unit-level outputs. The *reported* interval is the model's own. The *rescaled* interval divides the unit-level variances by m_S, the standard correction for miscalibrated unit-level intervals, and then aggregates at the unit level as before. The + *design effect* interval additionally multiplies the aggregate variance by $1 + (k - 1)\rho_S$. The CNP row averages its three seeds, each corrected with its own ρ_S and m_S.

In the three CausalPFN rows $m_S > 1$, and rescaling lowers coverage at both the smallest subset and $k_{\max}$. Under unit-level aggregation, only over-wide unit-level intervals offset the design effect, and rescaling removes them without correcting the aggregation. The CNP's unit-level intervals are close to calibrated ($m_S = 0.93$), so rescaling changes little and coverage at $k_{\max}$ stays near 0.24. CausalPFN's own temperature calibration acts like rescaling. Its full-sample interval on ACIC 2016 covers the true value in 18 of 40 datasets without it and in 8 with it.

The design-effect interval removes the dependence on k, as Equation (9) implies, but it requires ρ_S and m_S measured against the true effects and so cannot be used in practice.

D Subgroup selection

In Table 3, subsets are selected on the propensity score or on covariates, as an analyst querying a subgroup would select them, instead of uniformly at random.

Narrowing a pool raises ρ_S but shrinks $k_{\max}$, so the largest reachable design effect changes much less than either and falls for the narrowest pools.

Narrow propensity strata also lower m_S, whereas nested covariate cells leave it roughly unchanged and covariate neighbourhoods raise it. A rule that raises ρ_S without lowering m_S need not reduce coverage below nominal.

At $s = 1$, nested covariate cells reach ρ_S comparable to the bottom 25% propensity stratum and still never fall below nominal, because m_S stays high. At $s = 4$, where m_S starts lower, the first two covariate cells do.

E How existing models construct subset-CATE intervals

CausalPFN. Each unit's predictive distribution is a histogram over 1024 bins (Balazadeh et al., 2025a). The subset-CATE point estimate is the average of the relevant unit-level point estimates. For

subset	ρ_S	m_S	$k_{\max}$	$1+(k_{\max}-1)\rho_S$	$\hat{k}^*$
$s=1$, *propensity stratum*					
all units (baseline)	0.034	21.90	500	17.9	398
bottom 50%	0.069	23.23	250	18.3	216
bottom 25%	0.120	14.10	125	15.9	83
bottom 10%	0.224	8.07	44	10.6	34
$s=1$, *nested covariate cell*					
$X_1 > 0$	0.043	22.89	177	8.7	—
$\wedge\, X_2 > 0$	0.076	23.83	89	7.7	—
$\wedge\, X_3 > 0$	0.122	26.46	31	4.7	—
$s=1$, *covariate neighbourhood*					
nearest 50%	0.075	45.46	250	19.6	—
nearest 25%	0.143	46.08	125	18.7	—
nearest 10%	0.231	36.09	44	10.9	—
$s=4$, *propensity stratum*					
all units (baseline)	0.059	9.20	500	30.4	93
bottom 50%	0.076	9.72	250	20.0	89
bottom 25%	0.142	6.12	125	18.6	31
bottom 10%	0.257	3.63	44	12.1	8
$s=4$, *nested covariate cell*					
$X_1 > 0$	0.069	9.41	177	13.1	86
$\wedge\, X_2 > 0$	0.095	9.58	89	9.4	58
$\wedge\, X_3 > 0$	0.176	8.44	31	6.3	—

Table 3: Subset selection rules on CausalPFN at $n = 500$, with intervals formed by unit-level aggregation, over 64 tasks with 48 subsets per task. $k_{\max}$ is the largest evaluated subset size within the pool and $1+(k_{\max}-1)\rho_S$ the design effect at that size; "—" means coverage never falls below nominal within the pool. $\hat{k}^*$ is defined in Section 5.1.

the credible interval, its code draws one sample from each unit's predictive distribution, averages those samples across the units in S, repeats this many times, and takes percentiles of the resulting distribution of averages. For every subset, including the full sample, we apply this procedure to per-unit draws from CausalPFN's sampler.

Let $\hat{\tau}_i^{(b)}$ denote the b-th draw from unit i's predictive distribution, and let $\hat{\theta}^{(b)} = k^{-1}\sum_{i\in S}\hat{\tau}_i^{(b)}$ be the corresponding draw of the subset-CATE. Given $\mathcal{D}$, draws for different units are generated independently and share no randomness, so the covariance terms vanish and the variance of each aggregate draw is exactly

$$\mathrm{Var}\left(\hat{\theta}^{(b)}\right) = \frac{1}{k^2}\sum_{i\in S}\mathrm{Var}\left(\hat{\tau}_i^{(b)}\right) + \frac{1}{k^2}\sum_{\substack{i,j\in S\\ i\neq j}}\mathrm{Cov}\left(\hat{\tau}_i^{(b)},\hat{\tau}_j^{(b)}\right) = \frac{1}{k^2}\sum_{i\in S}\hat{\sigma}_i^2 \;=\; \hat{\sigma}_S^2, \tag{24}$$

where $\hat{\sigma}_i^2$ is the variance of unit i's predictive distribution. This is Equation (7), and the percentiles the code returns are Monte Carlo estimates of the quantiles of this distribution.

The interval the code returns is a percentile interval of the sampled aggregate distribution instead of the symmetric interval $\hat{\theta}_S \pm z_{1-\alpha/2}\hat{\sigma}_S$. The two coincide when the sampled distribution is symmetric, and the average of k independent draws approaches a Gaussian as k grows, so the discrepancy is largest at small k and negligible at the subset sizes where coverage fails. Whether $\hat{\sigma}_i^2$ matches unit i's actual squared error is a separate question, which m_S measures.

Do-PFN and CausalFM. Neither provides an aggregate estimator. The only uncertainty quantification in Do-PFN's repository, which is not part of its estimation interface, builds an interval from the standard error across units and so also assumes independence. CausalFM returns a Gaussian mixture for each unit, and the interval construction in its documentation samples each unit's mixture independently and forms a separate interval for each unit. The only aggregate quantity in its documentation is an average of unit-level point estimates, with no associated credible interval.

benchmark	n	ρ_S	m_S	**nRMSE**	**cov at k=n**	**se**	$\hat{k}^*$ ($\hat{k}^*/n$)
Hill-A	1000	0.0771	35.13	0.07	0.82	0.06	502 (50%)
Friedman	1000	0.0334	11.75	0.10	0.70	0.07	383 (38%)
IHDP	747	0.0420	5.43	0.11	0.60	0.08	85 (11%)
Hill-B heavy-tailed	1000	0.0077	2.03	0.30	0.53	0.08	62 (6.2%)
Hill-B	1000	0.0038	1.42	0.28	0.53	0.08	58 (5.8%)
Kang–Schafer	1000	0.0185	1.43	0.09	0.30	0.07	21 (2.1%)
ACIC 2016	4802	0.0202	1.63	0.27	0.20	0.06	30 (0.6%)

Table 4: The benchmarks of Figure 1 and the heavy-tailed variant of Hill-B, which the figure omits: CausalPFN with its temperature calibration, unit-level aggregation, subsets drawn uniformly at random, 40 datasets per benchmark. $\hat{k}^*$ is the subset size at which observed coverage first falls below nominal, interpolated in $\log k$, with $\hat{k}^*/n$ as a percentage of the sample. nRMSE is the root-mean-square error of the per-unit effect estimates divided by the standard deviation of the observed outcome, within each dataset and averaged over datasets. Rows are ordered by $\hat{k}^*/n$.

model	**aggregation**	**k=5**	**k=50**	**k=500**	**k=2000**
CNP	unit-level	0.93	0.61	0.29	0.17
latent NP	SCM-marginalized	1.00	0.98	0.97	0.97
CNP	SCM-marginalized	0.95	0.98	0.98	0.98

Table 5: Coverage of nominal 95% subset-CATE intervals at $n = 2000$ over 96 tasks with 24 subsets each. The two CNP rows are separately trained models. Each model was trained once, with optimization settings that differ from Appendix B and from each other, so the comparison is not controlled for optimization.

n	ρ_S	m_S	$\hat{k}^*$	$1 + (n-1)\rho_S$	**cov at k=n**
125	0.0385	7.03	—	6	0.95 ± 0.02
250	0.0422	14.25	—	12	0.96 ± 0.02
500	0.0339	21.90	398	18	0.94 ± 0.03
1000	0.0230	32.67	980	24	0.95 ± 0.02
2000	0.0230	40.08	1376	47	0.93 ± 0.03

Table 6: CausalPFN on the synthetic family of Appendix A at five sample sizes, with subsets drawn uniformly at random, 64 tasks and 48 subsets per task; coverage at $k = n$ is given $\pm$ one standard error clustered by task. The same CausalPFN model is used throughout, so no training choice varies down the table. $1 + (n-1)\rho_S$ is the design effect at the full sample; "—" means coverage never falls below nominal. $\hat{k}^*$ is defined in Section 5.1.

F ADDITIONAL RESULTS

Sample size. Whether ρ_S falls with n depends on the prior. When a few shared parameters drive the effect, as in our synthetic family, errors on different units stay correlated as the posterior concentrates, so ρ_S need not fall at all. If it fell as n^{-1}, the design effect at a fixed fraction of the sample would stay constant in n. Table 6 gives results for CausalPFN on the synthetic family of Appendix A at five sample sizes.

The decay is far slower than n^{-1}. A power law fitted to the five points has exponent -0.24, and over the sixteenfold range of n in the table ρ_S falls by a factor of 1.7. The largest subset grows with n while the correlation barely falls, so the design effect at the full sample, $1 + (n-1)\rho_S$, grows from 6 at $n = 125$ to 47 at $n = 2000$.

Coverage at $k = n$ nonetheless stays close to nominal across the table, because m_S grows as well. CausalPFN's per-unit errors shrink faster than the intervals it reports for them, so m_S rises from 7 to 40, with a fitted exponent of 0.62, and the numerator of Equation (9) grows about as fast as the denominator. On the benchmarks of Section 5.2, m_S lies between 1.42 and 35.1, and coverage falls below nominal on every one of them.

Accuracy of the predictions. The measured $\hat{k}^*$ is within 15% of Equation (11) on Hill-A, Kang–Schafer, and ACIC 2016, 19% larger on Friedman, and 20% smaller on IHDP. On Hill-B and its heavy-tailed variant, it is about half the prediction, and Equation (1) overstates full-sample coverage there (0.72 and 0.66, against 0.53) and on Kang–Schafer (0.41, against 0.30). One possible reason is that the predictions use moments pooled over the 40 datasets. Coverage is concave in R, so variation in R across datasets can lower pooled coverage below the prediction (Appendix G). On the synthetic family (Table 6), coverage stays within sampling error of nominal, so $\hat{k}^*$ is noisy there.

The heavy-tailed variant. Hill-B heavy-tailed is omitted from Figure 1 for legibility and appears in Table 4. Replacing Hill-B's Gaussian noise with t_3 noise changes little. Coverage falls below nominal at 6.2% of the sample, against 5.8% for Hill-B, and both reach 0.53 at the full sample.

G PROOFS

Throughout, $\mathbb{E}$, Var, Cov, and Pr are over $\zeta \sim \pi$, $\mathcal{D} \mid \zeta$, and S, as in Equation (5). All errors have finite second moments, and $b_i := \mathbb{E}[e_i \mid \mathcal{D}]$. We say that S is *selected on the data* if S is conditionally independent of $(\tau(X_i;\zeta))_{i=1}^n$ given $\mathcal{D}$. Uniformly drawn subsets and subsets chosen by covariates or by the model's own estimates are selected on the data. In our synthetic family, subsets chosen by the true propensity score are as well, because w_p is independent of (w_b, w_e) under the prior and enters a separate factor of the likelihood, so the two remain independent given $\mathcal{D}$. Subsets chosen by the true effect are not.

Proof of Proposition 1. Since $\bar{e}_S = k^{-1}\sum_{i\in S} e_i$,

$$k^2\,\bar{e}_S^2 \;=\; \sum_{i\in S} e_i^2 \;+\; \sum_{\substack{i,j\in S\\ i\neq j}} e_i e_j.$$

Taking expectations and using Equation (6) gives $k^2\mathbb{E}[\bar{e}_S^2] = k\sigma_S^2 + k(k-1)c_S$, which is Equation (8). By Equation (7), $\mathbb{E}[\hat{\sigma}_S^2] = k^{-1}\mathbb{E}[\tilde{\sigma}_S^2] = k^{-1}m_S\sigma_S^2$, and dividing by Equation (8) gives Equation (9). If $\hat{\sigma}_S^2$ takes the same value v_k for every query of size k, then $v_k = \mathbb{E}[\hat{\sigma}_S^2] = R\,\mathbb{E}[\bar{e}_S^2]$, and $\theta_S \in \hat{\theta}_S \pm z_{1-\alpha/2}\sqrt{v_k}$ if and only if $|\bar{e}_S| \le z_{1-\alpha/2}\sqrt{R\,\mathbb{E}[\bar{e}_S^2]}$, which is Equation (10). If in addition $\bar{e}_S$ is mean-zero Gaussian, the standardized error is standard normal and $\mathcal{C}(k) = 2\Phi(z_{1-\alpha/2}\sqrt{R}) - 1$. □

The same expansion, with $\mathbb{E}[(\sum_{i\in S} e_i)^2] \ge 0$ and $(\sum_{i\in S} e_i)^2 \le k\sum_{i\in S} e_i^2$, gives $-1/(k-1) \le \rho_S \le 1$ for $k \ge 2$.

Coverage when the reported variance varies. Suppose that, given $(\mathcal{D}, S)$, $\bar{e}_S$ is Gaussian with mean zero and variance $v_{\mathcal{D},S}$, as for the posterior mean when S is selected on the data and the posterior of θ_S is Gaussian. Then $\Pr(\theta_S \in [\hat{l}_S, \hat{u}_S] \mid \mathcal{D}, S) = f(R_{\mathcal{D},S})$ with $f(r) = 2\Phi(z_{1-\alpha/2}\sqrt{r}) - 1$ and $R_{\mathcal{D},S} = \hat{\sigma}_S^2 / v_{\mathcal{D},S}$, so $\mathcal{C}(k) = \mathbb{E}[f(R_{\mathcal{D},S})]$. The ratio R of Equation (9) is the average of $R_{\mathcal{D},S}$ weighted by $v_{\mathcal{D},S}/\mathbb{E}[v_{\mathcal{D},S}]$. Since f is concave, $\mathcal{C}(k) \le f(R)$ when $R_{\mathcal{D},S}$ is independent of $v_{\mathcal{D},S}$, and Equation (1) then overstates coverage. Without this independence the direction can reverse. If $\hat{\sigma}_S$ is fixed and only $v_{\mathcal{D},S}$ varies, coverage is convex in $v_{\mathcal{D},S}$ wherever it is below $2\Phi(\sqrt{3}) - 1 \approx 0.92$, and pooling then raises it above $f(R)$.

Proof of Proposition 2. Given $\mathcal{D}$, each $\hat{\tau}_i$ and each X_i is fixed, and $\tau_i = \tau(X_i;\zeta)$ with $\zeta \sim \pi(\zeta \mid \mathcal{D})$. Since e_i differs from $-\tau_i$ by a quantity that is fixed given $\mathcal{D}$, $\mathrm{Cov}(e_i, e_j \mid \mathcal{D}) = \mathrm{Cov}(\tau_i, \tau_j \mid \mathcal{D})$, which is Equation (13). If S is selected on the data, the law of the unit effects given $(\mathcal{D}, S)$ equals their law given $\mathcal{D}$, so for $i, j \in S$,

$$\mathbb{E}[e_i e_j \mid \mathcal{D}, S] \;=\; \mathrm{Cov}_{\zeta\sim\pi(\zeta\mid\mathcal{D})}\big(\tau(X_i;\zeta), \tau(X_j;\zeta)\big) + b_i b_j.$$

Averaging over pairs of distinct units in S and then over $(\mathcal{D}, S)$ gives c_S. The posterior mean under π has $b_i = 0$. □

Aggregation for a general model. For a model with joint predictive distribution $Q_S(\cdot \mid \mathcal{D})$ for $(\tau_i)_{i\in S}$ and marginals $Q_i(\cdot \mid \mathcal{D})$, unit-level aggregation reports the law of $k^{-1}\sum_{i\in S}\tilde{\tau}_i$ under

$(\tilde{\tau}_i)_{i\in S} \sim \bigotimes_{i\in S} Q_i$, and SCM-marginalized aggregation reports its law under $(\tilde{\tau}_i)_{i\in S} \sim Q_S$. For the exact posterior, Q_S is the law of $(\tau(X_i;\zeta))_{i\in S}$ under $\zeta \sim \pi(\zeta \mid \mathcal{D})$, and the two are Equations (2) and (3) with $p(\tau_i \mid \zeta)$ the point mass at $\tau(X_i;\zeta)$. For the latent NP, Q_S is the distribution in Equation (17). The CNP outputs a Gaussian for θ_S directly (Proposition 3), and a model that returns only marginals, such as CausalPFN, supports only unit-level aggregation.

Corollary 1 (Exact posterior)**.** *Let $\hat{\tau}_i = \mathbb{E}[\tau_i \mid \mathcal{D}]$ and $\hat{\sigma}_i^2 = \mathrm{Var}(\tau_i \mid \mathcal{D})$ under π, and let S be selected on the data. Then (i) unit-level aggregation has $m_S = 1$ and $R = 1/(1 + (k-1)\rho_S)$; (ii) SCM-marginalized aggregation, which reports $\mathrm{Var}(\theta_S \mid \mathcal{D})$, has $R = 1$ with this variance in place of $\hat{\sigma}_S^2$; and (iii) if the posterior of θ_S has a continuous distribution function, its equal-tailed $1-\alpha$ credible interval has $\mathcal{C}(k) = 1-\alpha$ for every k.*

Proof. (i) For $i \in S$, $\mathbb{E}[e_i^2 \mid \mathcal{D}, S] = \mathrm{Var}(\tau_i \mid \mathcal{D}) = \hat{\sigma}_i^2$, so $\mathbb{E}[\bar{\hat{\sigma}}_S^2] = \sigma_S^2$ and $m_S = 1$, and Proposition 1 gives R. (ii) The point estimate is $\mathbb{E}[\theta_S \mid \mathcal{D}]$ and the reported variance is $\mathrm{Var}(\theta_S \mid \mathcal{D})$, so $\mathbb{E}[\bar{e}_S^2] = \mathbb{E}[\mathrm{Var}(\theta_S \mid \mathcal{D})]$, which equals the expected reported variance. (iii) Given $(\mathcal{D}, S)$, θ_S follows its posterior, so the interval contains it with conditional probability $1-\alpha$. Averaging over $(\mathcal{D}, S)$ gives the claim. □

Part (iii) concerns coverage averaged over π, as in Equation (5), and does not imply coverage at a fixed SCM. In our synthetic family the outcome is linear in Gaussian coefficients with known noise variance, and w_p enters a separate factor of the likelihood, so the posterior of $(\tau_i)_{i\in S}$ is Gaussian and the interval in (iii) is $\hat{\theta}_S \pm z_{1-\alpha/2}\,\mathrm{Var}(\theta_S \mid \mathcal{D})^{1/2}$. Because s enters only the propensity factor, this posterior is the same function of $\mathcal{D}$ at every s, so the corollary applies at $s = 4$ as well, and the population minimizer of Proposition 3 at $s = 1$ is also the exact posterior at $s = 4$.

Marginals do not determine the joint. Let the unit effects have marginals $\mathcal{N}(\mu_i, \omega^2)$, $i \in S$. For every $\rho \in [0,1]$, the Gaussian with these marginals and all pairwise correlations ρ gives $\mathrm{Var}(k^{-1}\sum_{i\in S}\tau_i) = \omega^2\{1 + (k-1)\rho\}/k$. This ranges from ω^2/k under independence to ω^2 at $\rho = 1$, the largest value any joint distribution with these marginals allows, since the standard deviation of a sum is at most the sum of the standard deviations. The marginals are thus consistent with every design effect between 1 and k, and no function of them alone equals $\mathrm{Var}(\theta_S \mid \mathcal{D})$ for every joint posterior. A rule that also uses the model's prior can sometimes do better. In a linear-Gaussian model with known features, for instance, the marginal variances at enough units determine the posterior covariance of the coefficients. The statement in Section 4.2 concerns rules that use the marginals alone.

Proposition 3 (Bayes-optimal Gaussian output)**.** *Let training tasks draw $\zeta \sim \pi_0$, $\mathcal{D} \mid \zeta$, and a subset S selected on the data, with $\mathbb{E}_{\pi_0}[\theta_S^2] < \infty$ and $\mathbb{E}_{\pi_0}|\log \mathrm{Var}_{\pi_0}(\theta_S \mid \mathcal{D})| < \infty$. Among all measurable maps $(\mathcal{D}, S) \mapsto (\mu, v)$, the expected loss $\mathbb{E}_{\pi_0}\big[-\log \mathcal{N}\big(\theta_S; \mu(\mathcal{D}, S), v(\mathcal{D}, S)\big)\big]$ is minimized by $\mu = \mathbb{E}_{\pi_0}[\theta_S \mid \mathcal{D}]$ and $v = \mathrm{Var}_{\pi_0}(\theta_S \mid \mathcal{D})$, uniquely up to sets of probability zero under the training distribution of $(\mathcal{D}, S)$.*

Proof. Given $(\mathcal{D}, S)$, the conditional expected loss equals $\frac{1}{2}\log v + \{\mathrm{Var}(\theta_S \mid \mathcal{D}, S) + (\mathbb{E}[\theta_S \mid \mathcal{D}, S] - \mu)^2\}/(2v)$ plus a constant, which is uniquely minimized at $\mu = \mathbb{E}[\theta_S \mid \mathcal{D}, S]$ and $v = \mathrm{Var}(\theta_S \mid \mathcal{D}, S)$. Because S is selected on the data, these equal the posterior mean and variance of θ_S given $\mathcal{D}$. Minimizing the conditional loss for each $(\mathcal{D}, S)$ minimizes its expectation. □

As in Müller et al. (2022, Insight 1), the training loss is an expected cross-entropy to the posterior, and restricted to Gaussian outputs its minimizer matches the posterior's first two moments. The result describes the population minimizer. A trained network approximates it well only for tasks like those seen in training, which for our CNP include subsets of every size but not datasets at $s = 4$.