SAMPLE-SIZE EXTRAPOLATION IN SIMULATION-BASED INFERENCE VIA STATISTICAL SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

In simulation-based inference (SBI), neural posterior estimators trained on a bounded range of sample sizes can become inaccurate on larger datasets because their parameterizations do not specify how posterior location, uncertainty and shape must change as observations accumulate. We propose center-estimated NPE (CE-NPE), which combines a statistical center with an explicit contraction rate and learns a local center correction and residual densities in posterior-scale coordinates. The center is a statistical estimate, such as a closed-form statistic or a simulator-based moment match, which does not require an evaluable likelihood. Our analysis explains how contraction amplifies center error and decomposes Gaussian posterior error into shape, mean and covariance contributions. On six regular benchmarks, CE-NPE gives the best reference-posterior agreement of all tested amortized estimators at every out-of-range size, up to five times the largest training size, without retraining; it also fits better within the training range on most of them. Matched controls and nonregular examples characterize the scope of these gains and identify the conditions on which they depend: an appropriate contraction rate, an accurate center, and a stable residual representation.

1 INTRODUCTION

Simulation-based inference (SBI) targets models whose likelihood cannot be evaluated but which can be simulated (Cranmer et al., 2020). Neural posterior estimation (NPE), its most direct form, amortizes Bayesian inference by learning a conditional density from simulations and reusing it on new datasets (Papamakarios & Murray, 2016; Greenberg et al., 2019; Lueckmann et al., 2021). In deployment, however, the estimator may be asked to condition on n observations when n exceeds every sample size seen during training. Permutation-invariant encoders such as Deep Sets accommodate this change in cardinality (Zaheer et al., 2017), but the posterior must also change its location, width and shape as evidence accumulates, and providing the network with n as an input (Radev et al., 2022; Habermann et al., 2025; Mittal et al., 2025) leaves these relationships to be learned over a bounded training range. Accepting arbitrary dataset sizes is therefore an architectural property, whereas extrapolating posterior uncertainty is a statistical one. We ask **how posterior estimators trained on small datasets can extend to larger sample sizes without retraining**.

Statistical asymptotics describe how a posterior should change with n, and therefore supply structure for this problem. Under regularity conditions, the Bernstein–von Mises theorem describes an approximately Gaussian posterior with width of order $n^{-1/2}$ after appropriate centering (van der Vaart, 1998, Ch. 10). More generally, a justified contraction rate $n^{-\alpha}$, which nonregular models can change (Bochkina & Green, 2014), suggests learning in local coordinates $n^{\alpha}(\theta - c_n)$. Working in these coordinates removes the need for the network to infer the leading contraction rate from the training range, but it makes center accuracy critical: a location error δ_n becomes a shift $n^{\alpha}\delta_n$ in those coordinates, so a small fixed bias can become large relative to posterior uncertainty. Moreover, even a center that contracts at the correct rate leaves errors in covariance and residual shape, so scaling alone does not ensure calibrated inference.

We propose center-estimated NPE (CE-NPE), which anchors the posterior at a statistical estimator $c_{\mathrm{stat}}(x_{1:n})$ and learns only a correction, multiplied by $n^{-\alpha}$. This design places the learned adjustment on the posterior's contraction scale and thereby reduces what must be extrapolated in parameter space; because it supplies structure that would otherwise have to be learned from simulations, it

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.

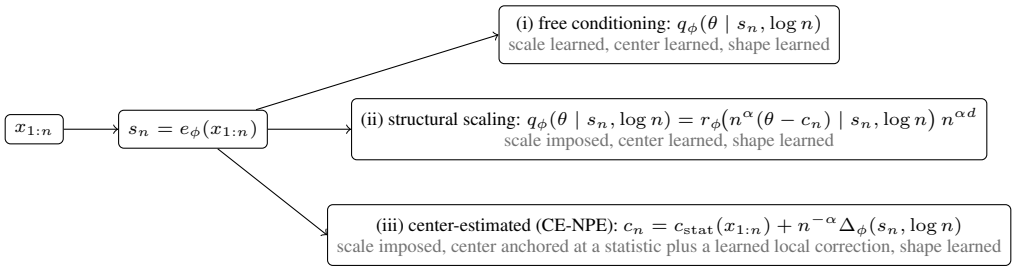

Figure 1: **Progressively removing what the network must extrapolate.** All three constructions share a permutation-invariant encoder and differ in which ingredients of the local coordinates $z_n = n^\alpha(\theta - c_n)$ are learned from the bounded training range and which are imposed: (i) free conditioning learns scale, center, and shape; (ii) structural scaling imposes the scale; (iii) CE-NPE additionally anchors the center at a statistical estimator.

can also improve fitting within the training range. The anchor can be a closed-form statistic or a simulator-based moment estimate, so it does not require an evaluable likelihood, and the network still learns uncertainty, dependence and residual shape. Figure 1 relates CE-NPE to free conditioning and to structural scaling with a learned center, which serves as an intermediate ablation.

Our experiments examine both the benefit of this parameterization and its source: six adapted regular benchmarks test reference-posterior agreement beyond the training range, matched-center controls assess what scaling adds beyond the statistical estimator, best-effort baselines assess sensitivity to training effort, and nonregular examples test the assumptions behind the local representation. Throughout, we report in-range performance alongside extrapolation results, separating errors already present within the training range from those that arise beyond it.

Contributions.

- **A structured parameterization**. We introduce CE-NPE, which combines a statistical center with an explicit contraction rate and learns the center correction and residual density in local coordinates. The center need not evaluate the likelihood and may be constructed from simple statistics or simulator access, so the estimator remains likelihood-free, and the parameterization reduces what the network must extrapolate beyond the training sample sizes (Section 3). The ingredients are classical; the contribution is to place them inside an amortized estimator and to show that its out-of-range gains are attributable to this structure rather than to network capacity.
- **Explaining the errors that remain after scaling.** We show how posterior contraction amplifies center error and separate Gaussian posterior error into shape, mean and covariance contributions. These results explain why correct scaling and a contracting center alone do not guarantee accurate, calibrated inference (Section 4).
- **Testing the benefits and limits of structural extrapolation.** We test whether explicit statistical structure improves amortized posterior estimation beyond the training sample-size range, under matched architectures and simulation budgets. On six regular benchmarks (Lueckmann et al., 2021), center anchoring gives the best reference agreement at every out-of-range size against free conditioning, structural scaling with a learned center, and a compositional score-based estimator (Geffner et al., 2023). Fixed-center controls isolate the contribution of scaling, while best-effort baselines assess sensitivity to training effort. Four nonregular prototypes with a faster rate, a boundary, a ridge and persistent multimodality show where the imposed structure fails, and a real-data application to 1000 Genomes windows (The 1000 Genomes Project Consortium, 2015) assesses consistency and held-out prediction without a reference posterior (Sections 5 and 6; Appendices D–G).

2 Related Work

Neural simulation-based inference. Neural SBI learns posterior densities, likelihoods, or likelihood ratios from simulations (Papamakarios & Murray, 2016; Greenberg et al., 2019; Papamakarios

et al., 2019; Hermans et al., 2020); see Cranmer et al. (2020) for a survey. Neural posterior estimation fits a conditional density estimator to simulated parameter–data pairs, either sequentially around one observation (Greenberg et al., 2019; Deistler et al., 2022) or amortized over the prior, and its density family has moved from normalizing flows (Durkan et al., 2019) to diffusion heads (Chen et al., 2025) and to transformers that serve many conditional queries with one model (Gloeckler et al., 2024). Amortization has also been extended to misspecified simulators (Ward et al., 2022) and beyond the standard Bayesian target, to generalized Bayesian posteriors indexed by a loss temperature (Sun et al., 2026). Benchmarks (Lueckmann et al., 2021) and calibration studies (Hermans et al., 2022) evaluate these estimators at fixed sample sizes. We study a different axis: how an amortized estimator trained on a bounded range of sample sizes transfers to larger datasets, and we introduce explicit statistical structure into its dependence on sample size.

Inference with variable-size datasets. Deep Sets and Set Transformer provide permutation-invariant representations of variable-size inputs (Zaheer et al., 2017; Lee et al., 2019), and Chan et al. (2018) apply exchangeable networks to likelihood-free inference from population genetic data. BayesFlow combines learned summaries with amortized posterior estimation, trains across variable numbers of observations, and shows posterior contraction empirically as observations accumulate (Radev et al., 2022). More recent amortized estimators condition transformers on sets of observations of varying size, within multilevel models (Habermann et al., 2025) and as in-context posterior estimation (Mittal et al., 2025). These architectures accommodate changing dataset size, but do not themselves prescribe a contraction rate. CE-NPE retains a set encoder and adds a specified rate and a statistical center to the posterior parameterization.

Compositional inference. For conditionally independent observations, learned likelihoods or likelihood ratios can be combined across observations to perform inference at different sample sizes (Papamakarios et al., 2019; Hermans et al., 2020). Compositional score methods similarly construct posterior samplers from networks trained on individual observations or small subsets (Geffner et al., 2023; Linhart et al., 2026). These methods exploit factorization to accommodate additional observations, but require the likelihood to factorize across subsets. CE-NPE instead models the full-data posterior in centered, rescaled coordinates, relying on an accurate statistical anchor and a stable residual representation rather than on a factorization (comparison with PF-NPSE in Appendix F.6).

Regression adjustments and learned summaries. ABC regression adjustment modifies simulated parameter draws to account for discrepancies between simulated and observed summaries (Beaumont et al., 2002), and semi-automatic ABC uses regression estimates of posterior means as summary statistics (Fearnhead & Prangle, 2012). CE-NPE uses an estimator for a different purpose, anchoring the coordinates in which an amortized posterior is learned, with a learned correction expressed in local coordinates and mapped to parameter space through $n^{-\alpha}$.

Asymptotic structure in posterior estimation. Under regularity conditions, the Bernstein–von Mises theorem motivates centering and rescaling a posterior by $\sqrt{n}$ (van der Vaart, 1998, Ch. 10). Asymptotic reasoning also informs parametric choices for amortized posterior approximation (Maceda et al., 2026), while Frazier et al. (2024) study the statistical accuracy of neural posterior and likelihood estimation. Our focus is transfer from a bounded training range in sample size: a justified contraction rate defines local coordinates, and we analyze the remaining errors from centering, covariance and residual shape.

3 METHODS

Let $\theta \in \Theta \subseteq \mathbb{R}^d$ have prior $p(\theta)$, and let $x_{1:n} \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta)$, where simulation from the generative model is available although the likelihood need not be evaluable. We train amortized estimators on simulated datasets with $n \leq n_{\max}$ and evaluate them at sample sizes $n > n_{\max}$. Every estimator is trained with the standard NPE objective, the log density assigned to the simulated parameter given the simulated dataset and its size (Papamakarios & Murray, 2016; Greenberg et al., 2019). Because the observations are exchangeable, all estimators share the same permutation-invariant Deep Sets encoder $s_n = e_\phi(x_{1:n})$ (Zaheer et al., 2017; Radev et al., 2022); the comparison focuses on how the

posterior parameterization depends on n, given a representation that already accepts variable-size datasets. The three constructions of Figure 1 differ only in how much of that dependence is learned and how much is imposed.

3.1 FREE CONDITIONING

The most direct approach is to provide sample size as an additional conditioning variable:

$$q_\phi^{\text{free}}(\theta \mid x_{1:n}, n) = q_\phi(\theta \mid s_n, \log n). \tag{1}$$

The logarithm only places the sample-size input on a convenient numerical scale. Apart from this explicit input, the construction is the standard one when a set encoder is combined with amortized posterior estimation (Zaheer et al., 2017; Radev et al., 2022). It imposes no constraint on how posterior location, scale, or shape evolve with n: any contraction must be inferred from the values of n seen in training, allowing flexible interpolation within the training range but requiring the learned dependence to be extended beyond it.

3.2 STRUCTURAL SCALING

In regular parametric problems the Bernstein–von Mises theorem implies that posterior uncertainty shrinks at the leading rate $n^{-1/2}$ (van der Vaart, 1998, Ch. 10). More generally, suppose the relevant contraction rate is $n^{-\alpha}$, with $\alpha = 1/2$ in the regular case, and define the rescaled variable

$$z_n = n^\alpha(\theta - c_n), \tag{2}$$

where c_n is a data-dependent center. In the basic structural estimator the center is learned from the encoded dataset, $c_n = c_\phi(s_n, \log n)$. Rather than modelling θ directly, we learn a conditional density r_ϕ for z_n:

$$z_n \mid x_{1:n} \sim r_\phi(\cdot \mid s_n, \log n). \tag{3}$$

The corresponding density in the original parameterization is

$$q_\phi^{\text{scale}}(\theta \mid x_{1:n}, n) = r_\phi(n^\alpha(\theta - c_n) \mid s_n, \log n)\, n^{\alpha d}. \tag{4}$$

Sample size therefore enters through the coordinate transformation and its Jacobian, as well as through the $\log n$ input to the network. The transformation imposes a leading contraction factor, while the network learns the center and residual distribution. Because a size-dependent residual can still change its own width, the explicit factor alone does not fix the fitted contraction rate.

3.3 CENTER-ESTIMATED NPE

With the rate imposed, the center c_n must still extrapolate, and in the rescaled coordinates a center error δ_n becomes $n^\alpha\delta_n$ (Section 4). CE-NPE therefore anchors the center at a problem-specific estimator $c_{\text{stat}}(x_{1:n})$ and lets the network $\Delta_\phi(s_n, \log n)$ learn only a correction:

$$c_n = c_{\text{stat}}(x_{1:n}) + n^{-\alpha}\Delta_\phi(s_n, \log n). \tag{5}$$

The factor $n^{-\alpha}$ expresses the learned correction in the same local scale as the posterior uncertainty; Proposition 1 (Appendix B) shows that with this factor a bounded correction preserves the anchor's error order, which a correction in parameter units does not. Substituting this center into $z_n = n^\alpha(\theta - c_n)$ gives

$$z_n = n^\alpha[\theta - c_{\text{stat}}(x_{1:n})] - \Delta_\phi(s_n, \log n). \tag{6}$$

The center is only a coordinate anchor: it need not be a MAP estimate, require an evaluable likelihood, or approximate posterior uncertainty or shape. Its key requirement is to track posterior location at the contraction rate, with error $O_p(n^{-\alpha})$, so that the error remains $O(1)$ in rescaled coordinates (Section 4). In our benchmarks, c_{stat} ranges from simple closed-form statistics such as $\bar{x}$ to EM or simulator-based moment matching (Appendix C).

3.4 RESIDUAL DENSITY FAMILIES

The structured parameterization does not require the rescaled posterior to be Gaussian, and we consider two residual families. The first is a Gaussian conditional density,

$$r_\phi(z_n \mid s_n, \log n) = \mathcal{N}(z_n; \mu_\phi(s_n, \log n), \Sigma_\phi(s_n, \log n)) \,, \tag{7}$$

which provides a simple residual family and lets us assess how far Gaussianity suffices after centering and scaling. This is not Gaussian variational inference: mean and covariance are amortized functions of $(s_n, \log n)$, and in parameter space the density has center $c_n + n^{-\alpha}\mu_\phi$ and covariance $n^{-2\alpha}\Sigma_\phi$. The second family uses a neural spline flow (NSF) (Durkan et al., 2019), which can represent non-Gaussian residual shapes. The free heads and the CE residual receive $\log n$ as a size input, so the conditional residual can account for finite-sample changes in shape and covariance, while the deterministic factor $n^{-\alpha}$ sets the contraction scale; the size inputs of the saved runs are documented in Appendix A.2.

4 WHY SCALING ALONE IS NOT ENOUGH

The factor $n^{-\alpha}$ fixes only the reference contraction scale; accurate extrapolation also requires that posterior location, uncertainty and shape remain correct as n grows.

4.1 KEEPING CENTER ERROR ON THE POSTERIOR SCALE

Let m_n be the true posterior mean and c_n the center of the residual coordinates. If posterior width decreases as $n^{-\alpha}$, a location error $c_n - m_n$ becomes a shift $n^\alpha(c_n - m_n)$ on the posterior scale: the same absolute error becomes five times larger in posterior-standard-deviation units when the posterior becomes five times narrower (Figure 2). The residual decomposes as

$$\underbrace{n^\alpha(\theta - c_n)}_{\text{residual to be learned}} = \underbrace{n^\alpha(\theta - m_n)}_{\text{variation around the true mean}} + \underbrace{n^\alpha(m_n - c_n)}_{\text{shift caused by center error}} \,, \tag{8}$$

Keeping this shift bounded requires the center error to decrease at least as fast as the posterior width. Otherwise, the residual network must learn a location shift that grows with sample size, even after the contraction scale has been removed.

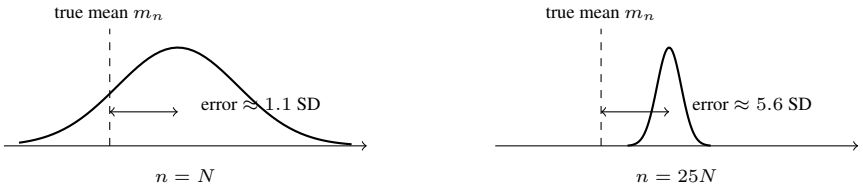

Figure 2: **The same location error becomes more serious as uncertainty shrinks.**

CE-NPE uses $c_n = c_{\text{stat}} + n^{-\alpha}\Delta_n$ to express the learned correction on this same scale (Equation 5). Proposition 1 (Appendix B) shows that, for a fixed trained network, if $c_{\text{stat}} - m_n = O_p(n^{-\alpha})$ and $\Delta_n = O_p(1)$ over repeated datasets, then $c_n - m_n = O_p(n^{-\alpha})$. A bounded local correction therefore preserves the anchor's error order; a correction added directly in parameter units must itself shrink at rate $n^{-\alpha}$ or faster to provide the same guarantee. Neither condition is enforced by training, and a bounded local correction cannot remove a persistent anchor bias; the anchor itself must therefore contract.

4.2 SEPARATING LOCATION, COVARIANCE, AND SHAPE ERRORS

Keeping the shift bounded is necessary but not sufficient: an error that remains bounded relative to posterior width need not be small. For example, two one-dimensional Gaussians with equal variances and means separated by one standard deviation have forward KL divergence $1/2$, even as both variances shrink.

Let P_n denote the true posterior, Q_n a fitted Gaussian posterior, and G_n the Gaussian with the same mean and covariance as P_n. Under the density, entropy, and covariance conditions stated in Proposition 3 (Appendix B.4),

$$\mathrm{KL}(P_n \| Q_n) = \underbrace{\mathrm{KL}(P_n \| G_n)}_{\text{non-Gaussian shape}} + \underbrace{\mathrm{KL}(G_n \| Q_n)}_{\text{mean and covariance mismatch}} . \tag{9}$$

The first term measures the error that remains even for the best Gaussian approximation (Proposition 3). The second measures errors in the fitted mean, marginal widths, and correlations. Because the learned residual can have a nonzero mean, the fitted mean need not equal c_n; controlling the center addresses only part of posterior accuracy.

For a fixed dataset, free conditioning, structural scaling, and statistical anchoring do not change the best achievable approximation within an unrestricted Gaussian family; they change how its parameters are learned and extrapolated. A spline residual can reduce the Gaussian shape error, but greater flexibility does not guarantee that training achieves this. The experiments therefore assess posterior-mean error, marginal width, and joint posterior agreement separately.

Both analyses hold at any fixed sample size; whether the fitted correction, covariance and residual shape remain accurate beyond the training range is an empirical question, which Section 5 addresses by reporting fit at the training boundary and beyond it separately.

5 Experiments

5.1 Experimental setup

Estimators and protocol. We compare five estimators under matched encoder widths, optimization schedules, and training simulation budgets: **Free-G** and **Free-NSF**, which learn the posterior through free conditioning; **Scale-G**, which imposes the contraction scale but learns the center; and **CE-G** and **CE-NSF**, which combine the imposed scale with a statistical anchor and a learned local correction. The suffixes G and NSF denote Gaussian and spline density heads, respectively. Training sizes are sampled uniformly from $n \in \{5, \ldots, 20\}$, and evaluation uses $n \in \{10, 20, 50, 100\}$, with datasets paired across estimators and seeds. Out-of-range sizes are used only for evaluation. All regular benchmarks impose the regular rate $\alpha = 1/2$. When no asymptotic argument supplies the rate, it can be estimated from how the anchor contracts across the training sizes; on our tasks this estimate serves as a check on the assumed rate rather than a substitute for it (Appendix F.4). Each training step draws a fresh batch of 128 simulated datasets for up to 4000 steps, evaluation uses 400 paired datasets per size, and reported values are means over three seeds (five on the mixture). Architectural and implementation details are given in Appendix A.

Tasks. We use six adapted regular benchmarks from the `sbibm` suite (Lueckmann et al., 2021): Gaussian linear with Gaussian and uniform priors, Gaussian mixture, Bernoulli GLM, SIR, and Lotka–Volterra. Each dataset contains n independent simulator replicates. The SIR and Lotka–Volterra priors are truncated to a regular regime in which the data remain informative about every parameter. Benchmark definitions, centers, and reference computations are given in Appendix C. Symmetric multimodal benchmarks and nonregular stress tests are reported in Appendix E.

Evaluation metrics. We assess posterior location, marginal width, and joint agreement with exact or numerically computed reference posteriors. Center error measures the fitted posterior-mean error in reference-standard-deviation units, averaged over coordinates; the SD ratio targets 1. Joint 95% HPD coverage (Hermans et al., 2022) targets 0.95, and C2ST (Lopez-Paz & Oquab, 2017) targets 0.5. These complementary diagnostics distinguish improvements in location and contraction from remaining errors in joint posterior structure (Section 4.2). KL divergence, MMD (Gretton et al., 2012) and simulation-based calibration (Talts et al., 2018) appear in the appendix studies that use them.

5.2 Does statistical structure improve sample-size extrapolation?

Overall performance. CE-G and CE-NSF achieve the lowest C2ST at both out-of-range sizes on all six regular benchmarks (Figure 3; full results at every size in Appendix D), and they also improve

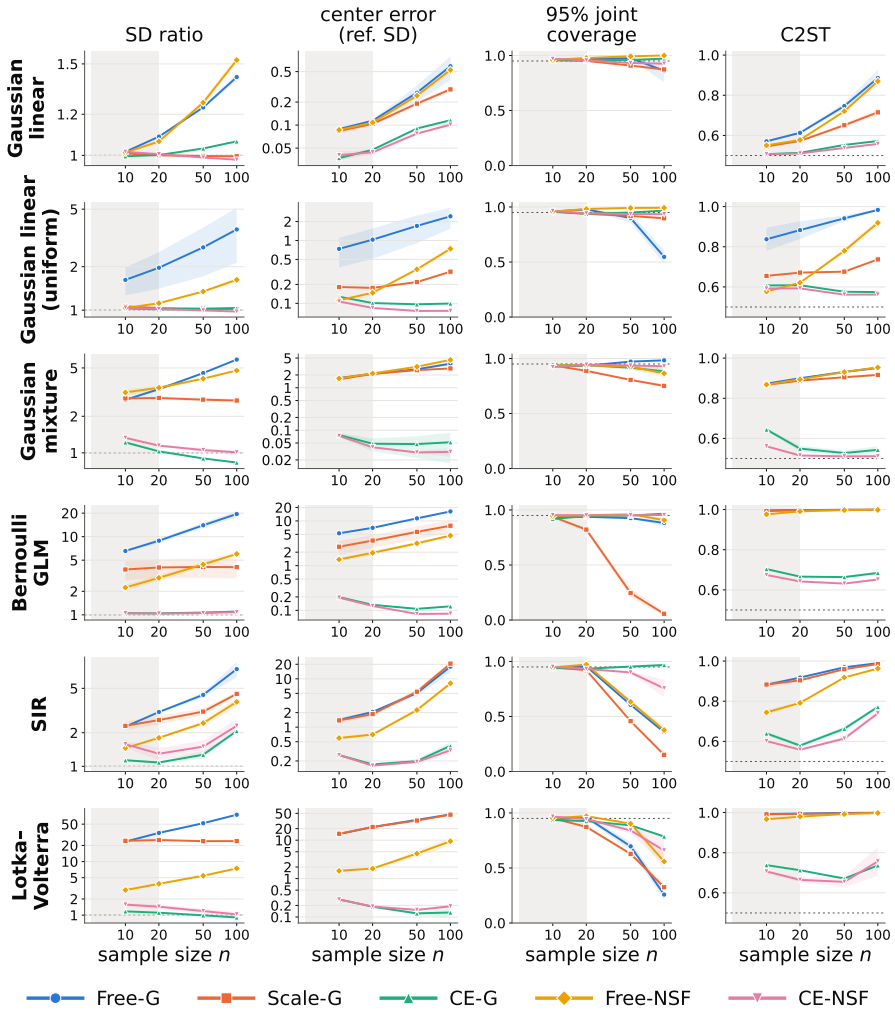

Figure 3: **Results on the six regular benchmarks.** Seed means with ± 1 seed-SD bands. The shaded band is the training range; dotted lines mark the targets (1, 0.95, and 0.5). SD ratio and center error are on log axes.

in-range agreement on most tasks. The in-range gain is expected: at n between 5 and 20 the posterior is already far narrower than the prior on these benchmarks, so the anchor and the imposed scale supply the location and width that a free network must otherwise learn. The advantage does not, however, extend uniformly to calibration: at $n = 100$, both CE estimators remain over-dispersed on SIR, where only CE-G maintains near-nominal joint coverage, and on Lotka–Volterra their marginal widths are close to the reference yet both under-cover.

Roles of scaling and centering. The intermediate ablation separates the two ingredients. Scale-G keeps its relative width approximately constant on five benchmarks and slows its growth on SIR, but because its center is still learned, its posterior-mean error in reference SD units increases with sample size on every benchmark. Anchoring the center removes this growth: CE-G improves both center error and C2ST over the Gaussian baselines at every out-of-range size. A fixed-center mixture control isolates the benefit of scaling more directly: with the same EM posterior mean, imposing the contraction factor substantially reduces C2ST and KL at $n = 100$, although coverage moves farther from nominal (Appendix F.3). This control also outperforms full CE-G in C2ST and KL, so on the mixture the gain comes from explicit scaling around an accurate center, and the learned location adjustment adds no further improvement. The same controls on SIR and Lotka–Volterra lead to the

same conclusion: the fixed-center, imposed-scale control matches or slightly improves on CE-G at every size, and the in-range advantage of CE-NSF narrows or reverses at $5\times$ (Appendix F.3).

Residual flexibility. The spline head adds a further, task-dependent benefit: CE-NSF improves on CE-G in width and C2ST on the mixture and matches it on the other benchmarks, whereas Free-NSF, which lacks the centered coordinates, remains inaccurate. Flexibility therefore helps once the coordinates are right, but does not substitute for them.

5.3 Real data: mutation rate and growth from 1000 Genomes

The regular benchmarks all have reference posteriors; population genetics provides a real-data application without one. Under a neutral coalescent model with growth, we infer the per-kb mutation rate and growth rate from genomic-window summaries, using a simulator-based moment estimator as the statistical anchor. The data are summaries of non-overlapping 10 kb windows on chromosome 22 of the 1000 Genomes phase-3 panel, and a dataset of size n is a subsample of n windows. The five estimators follow the main training-size protocol.

With no reference posterior, we use two complementary criteria (Figure 4): consistency with the moment fit to all windows, which shares the center's moment functional and therefore favors CE-NPE, and posterior-predictive energy distance to held-out windows, which is independent of that target.

As n grows to five times the training maximum, the posteriors of CE-G and CE-NSF contract onto the full-data moment fit, whereas Scale-G and Free-NSF contract onto a lower growth rate and Free-G barely contracts (Figure 4a); accordingly, only the CE estimators keep their consistency with the fit while the others drift away from it (Figure 4b). This stability costs nothing in predictive accuracy: on the anchor-independent criterion, their held-out predictions match the best free baseline at every size. In a separate in-model evaluation, CE has the lowest SBC KS statistics, although its joint coverage remains below nominal (Appendix G).

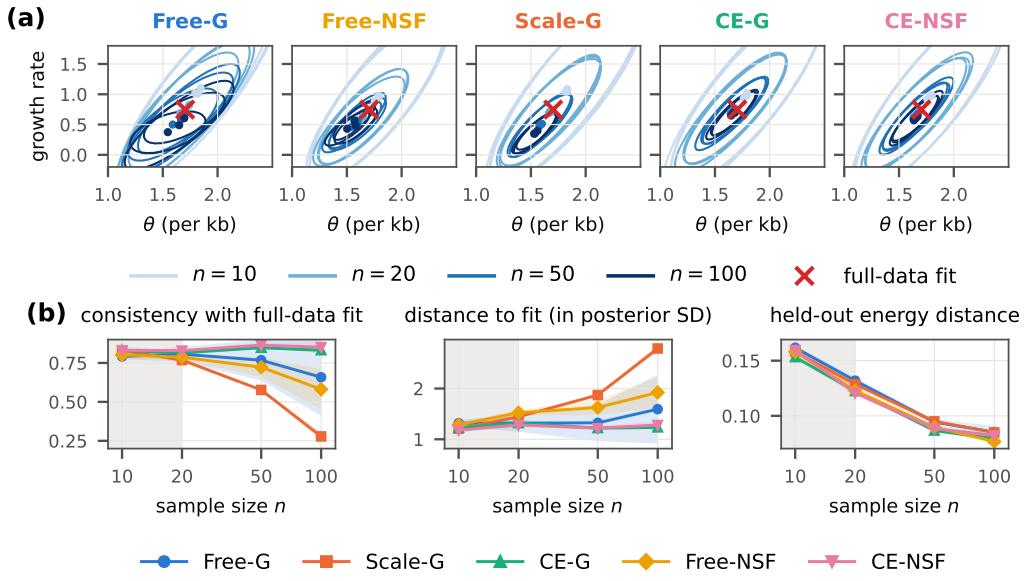

Figure 4: **Posterior contraction on real 1000 Genomes windows.** (a) The typical posterior of each estimator as n grows from 10 to 100, with the moment fit to all windows as the cross. (b) Against n: *consistency*, the fraction of datasets whose 95% intervals contain the full-data fit (higher is better); *distance*, how far the fit lies from the posterior mean in posterior SDs (lower is better); *energy distance* between posterior-predictive and held-out real windows (lower is better; independent of the anchor). Tables in Appendix G.

6 Discussion and limitations

Discussion. Taken together, the experiments support combining a statistical anchor with an explicit contraction scale: this structure reduces the sample-size dependence that a neural posterior estimator must learn, and it changes how posterior parameters are learned and extrapolated rather than enlarging the Gaussian approximation family (Section 4.2). The fixed-center controls clarify the source of these gains: with covariance still learned, anchoring at the statistic and imposing the scale match or improve on full CE-G in out-of-range C2ST on the mixture, SIR, and Lotka–Volterra (Appendix F.3). CE-NPE is accordingly intended for datasets comprising independent simulator replicates conditional on the parameter, when a contraction rate is justified and an affordable statistical center tracks posterior location on that scale.

Alternative explanations. Because several free baselines already fit poorly within the training range, we ran three controls to check that the out-of-range gap is not an artifact of under-training, of the particular training maximum, or of the choice of comparator (Appendix F). First, best-effort free baselines, retrained on every benchmark with a tuned learning rate, three times the step budget and longer patience, narrow the in-range gap on four benchmarks but leave the out-of-range ordering unchanged on all six. Second, retraining every estimator on SIR with a four times larger training maximum preserves CE's C2ST advantage at $5N$, although absolute accuracy and coverage change. Third, the compositional comparator PF-NPSE, trained on the same number of simulated datasets, approaches CE-NSF on Gaussian linear but remains less accurate on the other regular benchmarks, with several failures already present in range. Neither training effort, training range nor composition therefore accounts for the gap.

Scope and failure modes. The nonregular prototypes probe the rate, centering, and residual assumptions (Appendix E). On T1, whose posterior contracts at rate n^{-1} to a Gamma limit (Bochkina & Green, 2014), imposing the regular exponent produces increasing width mismatch; correcting the exponent substantially improves agreement (Table 11). T2 instead exposes a learning failure at a boundary: neither fitted residual recovers the shape, although a Gaussian renormalized on the parameter support can represent the exact posterior. A persistent ridge or separated modes, as in T3 and T4, do not become a stable local posterior under contraction about a single center. The two symmetric `sbibm` benchmarks, SLCP and two moons, behave in the same way: every single-center estimator fails at every size, including within the training range (Appendix E); a flexible flow can represent such distributions but must learn their changing geometry in the rescaled coordinates.

Practical and empirical limitations. CE-NPE is only partially amortized: simulator-based anchors can require numerical optimization at each query, adding task-specific computation beyond the neural estimator (Appendix C). The experiments cover finite simulation budgets and extrapolation up to five times the training maximum; additional training on every benchmark and a second training maximum on SIR test this sensitivity (Appendix F). Better agreement than the tested baselines does not by itself ensure calibrated credible regions, as the regular-benchmark and in-model population-genetic results show (Appendices D and G); calibration remains a separate concern in neural SBI (Hermans et al., 2022).

7 Conclusion

Accepting arbitrary dataset sizes is architectural; extrapolating posterior uncertainty is statistical. Explicitly encoding a justified posterior contraction rate removes the leading scale-extrapolation burden, a contracting statistical center controls the amplified location error, and once these first-order effects are stabilized, residual density learning targets the remaining covariance and shape errors. On six regular benchmarks CE-NPE gives the best reference agreement of the tested amortized estimators at every out-of-range size, up to five times the largest training size, without retraining; on real genomic windows CE-NPE stays consistent with the full-data fit at the same relative sizes at no cost in held-out predictive accuracy. Where the rate or the center is not justified, as in the nonregular prototypes, the same analysis identifies which ingredient fails and points toward representations adapted to the posterior geometry.

AI USE STATEMENT

LLM-based assistants helped polish the language and assisted with experiment, analysis, and figure code. The authors verified all content and take responsibility for the manuscript, code, and results.

REPRODUCIBILITY STATEMENT

Task-specific protocols, reference calculations and per-case results are provided in the appendix and accompanying code (supplementary material). Every table can be regenerated from the included per-case files. Where an evaluation uses a subset of the paired cases, the appendix states the subset.

REFERENCES

Mark A. Beaumont, Wenyang Zhang, and David J. Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

Natalia A. Bochkina and Peter J. Green. The Bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850 – 1878, 2014. doi: 10.1214/14-AOS1239. URL https://doi.org/10.1214/14-AOS1239.

Jeffrey Chan, Valerio Perrone, Jeffrey P. Spence, Paul A. Jenkins, Sara Mathieson, and Yun S. Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Tianyu Chen, Vansh Bansal, and James G. Scott. Conditional diffusions for amortized neural posterior estimation. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *PMLR*, pp. 2377–2385, 2025.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

Michael Deistler, Pedro J. Goncalves, and Jakob H. Macke. Truncated proposals for scalable and hassle-free simulation-based inference. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Conor Durkan, Iain Murray, and George Papamakarios. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, pp. 2771–2781. PMLR, 2020.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *arXiv preprint arXiv:1702.03118*, 2017. URL https://arxiv.org/abs/1702.03118.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: Semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B*, 74(3):419–474, 2012.

David T. Frazier, Ryan Kelly, Christopher Drovandi, and David J. Warne. The statistical accuracy of neural posterior and likelihood estimation. *arXiv preprint arXiv:2411.12068*, 2024. URL https://arxiv.org/abs/2411.12068.

Tomas Geffner, George Papamakarios, and Andriy Mnih. Compositional score modeling for simulation-based inference. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *PMLR*, pp. 11098–11116, 2023.

Manuel Gloeckler, Michael Deistler, Christian Weilbach, Frank Wood, and Jakob H. Macke. All-in-one simulation-based inference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Pedro J. Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F. Podlaski, Sara A. Haddad, Tim P. Vogels, David S. Greenberg, and Jakob H. Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:e56261, 2020.

David Greenberg, Marcel Nonnenmacher, and Jakob Macke. Automatic posterior transformation for likelihood-free inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pp. 2404–2414, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Daniel Habermann, Marvin Schmitt, Lars Kühmichel, Andreas Bulling, Stefan T. Radev, and Paul-Christian Bürkner. Amortized bayesian multilevel models. *Bayesian Analysis*, 2025. Advance publication, doi:10.1214/25-BA1570.

Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A crisis in simulation-based inference? beware, your posterior approximations can be unfaithful. *Transactions on Machine Learning Research*, 2022.

William Ogilvy Kermack and Anderson G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772):700–721, 1927.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

John Frank Charles Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248, 1982.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Julia Linhart, Gabriel Victorino Cardoso, Alexandre Gramfort, Sylvain Le Corff, and Pedro L. C. Rodrigues. Diffusion posterior sampling for simulation-based inference in tall data settings. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2404.07593.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1608.03983.

Alfred J. Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences*, 6(7):410–415, 1920.

Jan-Matthis Lueckmann, Pedro J. Goncalves, Giacomo Bassetto, Kaan Ö.cal, Marcel Nonnenmacher, and Jakob H. Macke. Flexible statistical inference for mechanistic models of neural dynamics. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *PMLR*, pp. 343–351, 2021.

Elliot Maceda, Emily C. Hector, Amanda Lenzi, and Brian J. Reich. Demonstrating the power and flexibility of variational assumptions for amortized neural posterior estimation in environmental applications. *Environmetrics*, 37(1):e70074, 2026. doi: 10.1002/env.70074.

Sarthak Mittal, Niels Leif Bracher, Guillaume Lajoie, Priyank Jaini, and Marcus Brubaker. Amortized in-context bayesian posterior estimation. *arXiv preprint arXiv:2502.06601*, 2025.

George Papamakarios and Iain Murray. Fast ϵ-free inference of simulation models with bayesian conditional density estimation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Stefan T. Radev, Ulf K. Mertens, Andreas Voss, Lynton Ardizzone, and Ullrich Köthe. Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(4):1452–1466, 2022.

Scott A. Sisson, Yanan Fan, and Mark M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

Shiyi Sun, Geoff K. Nicholls, and Jeong Eun Lee. Amortized simulation-based inference in generalized bayes via neural posterior estimation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526 (7571):68–74, 2015.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Daniel Ward, Patrick Cannon, Mark Beaumont, Matteo Fasiolo, and Sebastian M. Schmon. Robust neural posterior estimation and statistical model criticism. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, 1975.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Appendix

Roadmap. Appendix A gives the architecture and training protocol; Appendix B the statements and proofs behind Section 4; Appendix C the benchmark and reference definitions; Appendix D the six regular benchmarks at every size; Appendix E the nonregular stress tests; Appendix F the controls (best-effort baselines, a second training maximum, fixed-center, exponent estimation, size context, a compositional comparator); Appendix G the real-data application.

A Implementation and training protocol

The five estimators share the encoder, the optimization protocol and the training budget; they differ only in the density head and in how the center is obtained (Section 3).

A.1 Architecture

All five estimators use the same Deep Sets encoder (Zaheer et al., 2017):

- **Per-observation MLP:** $x_i \in \mathbb{R}^{d_x} \mapsto h_i \in \mathbb{R}^{64}$, with widths $d_x \to 64 \to 64 \to 64$.
- **Mean pooling:** $\bar{h} = n^{-1} \sum_{i=1}^{n} h_i$.
- **Aggregation MLP:** $[\bar{h}, t_n] \mapsto s_n \in \mathbb{R}^{64}$, with widths $(64 + d_t) \to 64 \to 64$.

Here $t_n \in \mathbb{R}^{d_t}$ contains the sample mean, log SD, skewness and excess kurtosis of each coordinate, plus task-specific summaries for SIR and Lotka–Volterra (Appendix C).

The density heads and center networks are:

- **Gaussian head:** MLP with hidden widths $(64, 64)$; outputs $\mu \in \mathbb{R}^d$ and a lower-triangular L, with $\Sigma = LL^\top$ and $L_{jj} = \exp(\mathrm{clip}(a_j, -7, 5))$.
- **NSF head** (Durkan et al., 2019): 5 transforms, 12 bins, conditioner hidden widths $(64, 64)$.
- **Center/correction:** MLP with hidden widths $(64, 64)$ and d outputs.

All MLPs use SiLU (Elfwing et al., 2017) between linear layers and no output activation.

Parameter coordinates. Free-G uses θ; Free-NSF uses $(\theta - \mu_{\text{prior}})/\sigma_{\text{prior}}$, with coordinate-wise moments estimated from 20,000 prior draws. Scale-G, CE-G and CE-NSF use $z_n = n^{\alpha}(\theta - c_n)$ (Equation 2) without further standardization.

A.2 Size inputs

We use $\log n$ to express neural size conditioning in the main-text parameterization. Sample size also enters through the explicit contraction factor $n^{-\alpha}$ and the statistical anchor computed from the observed dataset. The regular benchmarks use $\alpha = 1/2$.

Appendix F.5 compares $\log n$, $1/n$, and no size input to the residual network while holding the structured center and contraction factor fixed. On this Gamma diagnostic, the three choices give nearly identical contraction exponents and similar width and calibration diagnostics, with small differences in KL. This comparison concerns residual context; it does not establish equivalence of all size encodings or network components.

A.3 Training and evaluation

- Training: up to 4000 steps, each on a fresh batch of 128 simulated datasets, n drawn uniformly from $\{5, \ldots, 20\}$; Adam (Kingma & Ba, 2015) with learning rate 10^{-3}, cosine-annealed to zero over the run without restarts (Loshchilov & Hutter, 2017), and gradient norm clipped at 5. The objective is the negative log density of the simulated parameter under the estimator (Section 3).

- Validation: 192 fixed cases at each $n \in \{5, 10, 15, 20\}$ every 200 steps. Validation NLL is averaged across these sizes; the best checkpoint is retained. An improvement must exceed 10^{-4}, and training stops after 20 validation checks without such an improvement.
- Evaluation: 400 paired cases at $n \in \{10, 20, 50, 100\}$, identical across methods and seeds; 256 posterior draws for calibration metrics, 1024 for reference metrics (Section 5.1).
- Out-of-range sizes are never used for early stopping, hyperparameter selection or post-hoc calibration.
- Seeds: three on the regular benchmarks, except the mixture, which uses five for C2ST and coverage and two for the saved width and center-error summaries; the prototypes use ten. Values are means across seeds, with sample SDs where shown; table captions state exceptions.

The prototypes of Appendix E use the same training range and evaluation grid with a smaller network and their own budget, stated there. Two protocol variations are reported separately: best-effort free baselines (learning-rate sweep, three times the budget; Appendix F.1) and a second training maximum $N = 80$ on SIR (training $n \in \{20, \ldots, 80\}$, evaluation $n \in \{40, 80, 200, 400\}$; Appendix F.2).

B THEORETICAL SUPPORT

This appendix collects the statements and proofs behind Section 4. The first part concerns the center: what a contracting anchor means, why the learned correction is multiplied by the contraction factor, and how center and rate errors change with sample size. The second part concerns the fitted posterior as a whole: an exact decomposition of its error into shape, mean and covariance terms. Throughout, the network is fixed after training.

B.1 NOTATION AND THE MEANING OF A CONTRACTING CENTER

Let P_n be the true posterior for a dataset of size n, with mean m_n and covariance $n^{-2\alpha}\Sigma_n$, where $\Sigma_n \to \Sigma \succ 0$. The exponent $\alpha > 0$ specifies the contraction rate; Σ_n specifies the remaining covariance. All probability statements concern repeated datasets at a fixed generating parameter θ_0 after the network has been trained.

We distinguish three locations. The statistical anchor is $c_{\text{stat}}(x_{1:n})$. The residual coordinate is centered at $c_n = c_{\text{stat}} + n^{-\alpha}\Delta_n$, where Δ_n is the network's local correction. The mean of the final fitted posterior is a_n. If its residual has mean μ_n, then

$$a_n = c_{\text{stat}} + n^{-\alpha}(\Delta_n + \mu_n). \tag{10}$$

Thus, the anchor and the fitted posterior mean need not coincide. Only the sum of the two learned location terms affects a_n; an implementation may combine them into a single output.

The condition $n^{\alpha}(c_{\text{stat}} - m_n) = O_p(1)$ means that the anchor's error remains bounded in probability after multiplying by n^{α}. Because Σ_n has a positive-definite limit, this is equivalent to bounded error in posterior-standard-deviation units. It does not require that the standardized error tend to zero.

B.2 WHY MULTIPLY THE CORRECTION BY THE CONTRACTION FACTOR?

Proposition 1 (Scaling of center errors)**.** *Suppose $n^{\alpha}(c_{\text{stat}} - m_n) = O_p(1)$ and the fitted local correction satisfies $\Delta_n = O_p(1)$. Then*

$$c_n = c_{\text{stat}} + n^{-\alpha}\Delta_n \quad \Longrightarrow \quad n^{\alpha}(c_n - m_n) = O_p(1).$$

If instead the correction is applied in parameter units, then

$$c_n = c_{\text{stat}} + \Delta_n \quad \Longrightarrow \quad n^{\alpha}(c_n - m_n) = n^{\alpha}\Delta_n + O_p(1).$$

In particular, a correction converging to a nonzero constant makes the standardized center error diverge. A parameter-unit correction of order $O_p(n^{-\alpha})$ retains bounded error.

Proof. Subtract m_n and multiply each expression by n^{α}. For the local correction the result is $n^{\alpha}(c_{\text{stat}} - m_n) + \Delta_n$, a sum of two tight terms. For the parameter-unit correction the second term is $n^{\alpha}\Delta_n$, giving the stated alternatives. □

This is a sufficient-condition statement about the correction, not proof that a particular training procedure meets it. The result also requires a contracting anchor. If the anchor retains a fixed bias, a bounded local correction becomes too small in parameter units to cancel it. For the final mean a_n, the same argument applies to $\Delta_n + \mu_n$ in Equation 10.

B.3 HOW FAR CAN CENTER AND RATE ERRORS GROW?

Set $N = n_{\max}$ and write a deployment size as $n = fN$, with $f \geq 1$. The following calculation assumes specified power laws over the range of interest. The width σ_n is a scalar reference scale; it may describe one parameter direction. It is not a substitute for checking the full covariance in a multidimensional problem.

Proposition 2 (Center and rate sensitivity)**.** *Suppose $\|a_{fN} - m_{fN}\| \leq e_N f^{-\gamma}$ and $\sigma_{fN} = \sigma_N f^{-\alpha}$. Then*

$$\frac{\|a_{fN} - m_{fN}\|}{\sigma_{fN}} \leq \frac{e_N}{\sigma_N} f^{\alpha-\gamma}. \tag{11}$$

If the center-error law is an equality with $e_N > 0$ and $\alpha > \gamma$, a tolerance $\tau \geq e_N/\sigma_N$ is reached at

$$f_{\text{center}} = (\tau\sigma_N/e_N)^{1/(\alpha-\gamma)}.$$

Separately, suppose the fitted width is $\widehat{\sigma}_{fN} = \sigma_N f^{-\widehat{\alpha}}$, so that it matches the reference at N. With $\eta = \widehat{\alpha} - \alpha$,

$$\frac{\widehat{\sigma}_{fN}}{\sigma_{fN}} = f^{-\eta}, \qquad \left|\log \frac{\widehat{\sigma}_{fN}}{\sigma_{fN}}\right| \leq \kappa \iff |\eta| \log f \leq \kappa. \tag{12}$$

Proof. Divide the center-error bound by the reference width. Under equality, solve for f at the specified tolerance. For the width statement, divide the two power laws and take logarithms. □

Equation 11 has three cases: the bound on standardized error decreases if $\gamma > \alpha$, stays constant if $\gamma = \alpha$, and grows if $\gamma < \alpha$. These describe the error itself when the assumed center-error law is an equality; a growing upper bound alone does not establish that the error grows. With a fixed offset and root-n contraction, the error grows as $\sqrt{f}$ under equality. For example, increasing sample size from N to $25N$ multiplies the standardized error by five, even though the absolute offset has not changed. The threshold f_{center} is therefore conditional on the assumed power laws, not an extrapolation range guaranteed by training. Here the error concerns the final fitted mean a_n, including any residual mean, rather than the coordinate anchor alone. Equation 12 says that an exponent that is too large produces a posterior that becomes too narrow, and conversely for an exponent that is too small. The T1 prototype (Appendix E) is a direct test: its posterior contracts at rate n^{-1} while every estimator imposes $n^{-1/2}$, so $\eta = -1/2$ and the predicted over-dispersion is $\sqrt{f}$. The measured factors on T1 are 1.56 and 2.21 at $f = 2.5$ and $f = 5$, within 2% of the prediction for both residual families.

The fitted-width assumption is essential. A size-dependent residual can change its own width and partly undo an imposed exponent; the coordinate transformation alone does not imply that the fitted width follows the stated power law. These calculations also omit other sources of posterior error. They explain sensitivity under specified conditions, rather than establishing a deployment guarantee.

B.4 An exact decomposition for a Gaussian approximation

The true posterior P_n has mean m_n and covariance $n^{-2\alpha}\Sigma_n$. The fitted Gaussian Q_n has mean a_n and covariance $n^{-2\alpha}A_n$. The rate factor is shared in this notation, but A_n is allowed to depend on n, so the notation does not assume correct fitted contraction.

Proposition 3 (Exact error decomposition)**.** *Suppose P_n has a density with finite differential entropy and positive-definite covariance, and $A_n \succ 0$. Define*

$$G_n = \mathcal{N}(m_n, n^{-2\alpha}\Sigma_n), \qquad J_n = \mathrm{KL}(P_n \| G_n),$$

and set $h_n = n^\alpha(a_n - m_n)$ and $E_n = A_n^{-1/2}\Sigma_n A_n^{-1/2}$. Then

$$\mathrm{KL}(P_n\|Q_n) = \underbrace{J_n}_{\text{Gaussian shape error}} + \underbrace{\tfrac{1}{2}h_n^\top A_n^{-1} h_n}_{\text{mean error}} + \underbrace{\tfrac{1}{2}\{\mathrm{tr}(E_n) - d - \log\det E_n\}}_{\text{covariance error}}. \tag{13}$$

All three terms are nonnegative. If the eigenvalues of A_n and E_n stay in fixed compact subsets of $(0,\infty)$, there are constants $0 < c < C < \infty$ such that

$$c\{\|h_n\|^2 + \|E_n - I\|_F^2\} \leq \mathrm{KL}(P_n\|Q_n) - J_n \leq C\{\|h_n\|^2 + \|E_n - I\|_F^2\}.$$

Proof. Add and subtract $\log g_n$ inside $\mathbb{E}_{P_n}\log(p_n/q_n)$. The first term is J_n. The remaining expectation involves the log ratio of two Gaussian densities, which is quadratic in θ. Since P_n and G_n have the same first two moments, that expectation equals $\mathrm{KL}(G_n\|Q_n)$. The Gaussian KL formula gives Equation 13.

For the quadratic bounds, diagonalize E_n. Its covariance penalty is the sum of $g(\lambda) = \lambda - 1 - \log\lambda$ over its eigenvalues. Here $g(1) = g'(1) = 0$ and $g''(\lambda) = \lambda^{-2}$. Taylor's theorem on the stated compact interval bounds this sum above and below by constant multiples of $\|E_n - I\|_F^2$. The eigenvalue bounds for A_n give the corresponding bounds for the mean term. □

The first term is the error that remains even after choosing the best Gaussian mean and covariance. The second measures displacement in units of the fitted uncertainty. The third includes errors in widths, correlations and covariance orientation. None of these is removed merely by writing an explicit factor $n^{-2\alpha}$.

For CE-G, a_n includes both learned location terms in Equation 10; substituting the statistical anchor alone would omit part of the error. If P_n is itself Gaussian, $J_n = 0$. For a flow approximation, KL remains invariant under an invertible coordinate change, but the Gaussian three-term formula does not apply to its general density.

Bounded error can still be substantial. For $P_n = \mathcal{N}(0, 1/n)$ and $Q_n = \mathcal{N}(a/\sqrt{n}, 1/n)$, the formula gives $\mathrm{KL}(P_n \| Q_n) = a^2/2$ for every n. Both distributions contract at the same rate, yet their relative displacement stays fixed. Bounded standardized mean error therefore does not imply vanishing posterior error or nominal coverage.

The moment-matched Gaussian is the best Gaussian. Taking $a_n = m_n$ and $A_n = \Sigma_n$ in Equation 13 makes the mean and covariance terms vanish, and both are strictly positive otherwise, so G_n is the unique minimizer of $\mathrm{KL}(P_n \| \cdot)$ over Gaussians and J_n is the Gaussian family's irreducible error. Because an unrestricted Gaussian remains an unrestricted Gaussian after an invertible centering and scaling, Free-G, Scale-G and CE-G share this best achievable approximation at a fixed dataset; their fitted errors differ because they learn and extrapolate the mean and covariance differently. Only a more flexible residual can reduce J_n, and whether a trained spline does so is an empirical question (Section 5.2).

C BENCHMARK AND REFERENCE DEFINITIONS

C.1 GAUSSIAN LINEAR

Prior	$\theta \sim \mathcal{N}(0,\ 0.1\, I_{10})$
Simulator	$x_i \mid \theta \sim \mathcal{N}(\theta,\ 0.1\, I_{10})$
Dimensions	$\theta \in \mathbb{R}^{10}$, $x_i \in \mathbb{R}^{10}$
Center	sample mean $\bar{x}$
Reference posterior	conjugate Gaussian posterior
References	Lueckmann et al. (2021)

C.2 GAUSSIAN LINEAR (UNIFORM)

Prior	$\theta \sim U[-1, 1]^{10}$
Simulator	$x_i \mid \theta \sim \mathcal{N}(\theta,\ 0.1\, I_{10})$
Dimensions	$\theta \in \mathbb{R}^{10}$, $x_i \in \mathbb{R}^{10}$
Center	sample mean $\bar{x}$
Reference posterior	product of truncated normals
References	Lueckmann et al. (2021)

C.3 GAUSSIAN MIXTURE

Prior	$\theta \sim U[-10, 10]^2$
Simulator	$x_i \mid \theta \sim \frac{1}{2}\mathcal{N}(\theta, I) + \frac{1}{2}\mathcal{N}(\theta, 0.01\, I)$
Dimensions	$\theta \in \mathbb{R}^2$, $x_i \in \mathbb{R}^2$
Center	ten EM steps from the coordinate-wise median
Reference posterior	grid integration, 201×201 points around the MAP
References	Sisson et al. (2007); Beaumont et al. (2009)

C.4 BERNOULLI GLM

Inference of a ten-parameter generalized linear model with Bernoulli observations and a Gaussian prior whose covariance encourages smoothness by penalizing second-order differences in the filter; the observation is the sufficient statistic of the GLM.

Prior	$\theta = (\beta, f)$: $\beta \sim \mathcal{N}(0, 2)$, $f \sim \mathcal{N}\big(0, (F^\top F)^{-1}\big)$, with $F_{i,i-2} = 1$, $F_{i,i-1} = -2$, $F_{i,i} = 1 + \sqrt{(i-1)/9}$, $F_{i,j} = 0$ otherwise, $1 \le i, j \le 9$
Simulator	$x_i = (x_{i,1}, \ldots, x_{i,10})$, $x_{i,1} = \sum_{t=1}^{T} z_t$, $x_{i,2:10} = \frac{1}{x_{i,1}} V z$, $z_t \sim \mathrm{Bern}\big(\eta(v_t^\top f + \beta)\big)$, $\eta(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$, frozen input between time bins $t-8$ and t: $v_t \sim \mathcal{N}(0, I)$, $V = [v_1, \ldots, v_T]$
Dimensions	$\theta \in \mathbb{R}^{10}$, $x_i \in \mathbb{R}^{10}$
Fixed parameters	duration $T = 100$ time bins
Center	simulator method of moments: damped Gauss–Newton steps matching $X^\top \sigma(X\theta)$ to the replicate mean

Reference posterior adaptive importance sampling from the Newton MAP

References Lueckmann et al. (2017); Gonçalves et al. (2020)

C.5 SLCP

A task designed to have a simple likelihood and a complex posterior: one replicate is four two-dimensional points from a Gaussian whose mean and covariance are nonlinear functions of θ, so the posterior is a union of four mirror images.

Prior $\theta \sim U[-3, 3]^5$

Simulator $x_i = (x_{i,1}, \ldots, x_{i,4})$, $x_{i,j} \sim \mathcal{N}(m_\theta, S_\theta)$, where
$m_\theta = (\theta_1, \theta_2)^\top$, $S_\theta = \begin{pmatrix} s_1^2 & \rho s_1 s_2 \\ \rho s_1 s_2 & s_2^2 \end{pmatrix}$,
$s_1 = \theta_3^2$, $s_2 = \theta_4^2$, $\rho = \tanh \theta_5$

Dimensions $\theta \in \mathbb{R}^5$, $x_i \in \mathbb{R}^8$

Center moment estimator on the branch $\theta_3, \theta_4 > 0$

Reference posterior importance sampling on a fundamental domain

References Papamakarios et al. (2019); Greenberg et al. (2019); Hermans et al. (2020); Durkan et al. (2020)

C.6 Two moons

Prior $\theta \sim U[-1, 1]^2$

Simulator $x_i \mid \theta = \begin{bmatrix} r \cos \alpha + 0.25 \\ r \sin \alpha \end{bmatrix} + \begin{bmatrix} -|\theta_1 + \theta_2|/\sqrt{2} \\ (-\theta_1 + \theta_2)/\sqrt{2} \end{bmatrix}$, where $\alpha \sim U(-\pi/2, \pi/2)$, $r \sim \mathcal{N}(0.1,\ 0.01^2)$

Dimensions $\theta \in \mathbb{R}^2$, $x_i \in \mathbb{R}^2$

Center moment estimator on the branch $\theta_1 + \theta_2 \geq 0$

Reference posterior grid integration on a fundamental domain, expanded by the symmetry group

References Greenberg et al. (2019)

C.7 SIR

An epidemiological model of susceptible S, infectious I and recovered R individuals; the task infers the contact rate β and the recovery rate γ from binomial samples of I at ten evenly spaced times.

Prior $\beta \sim \mathrm{LogNormal}(\log 0.4,\ 0.5)$, $\gamma \sim \mathrm{LogNormal}(\log 0.125,\ 0.2)$,
truncated to $R_0 = \beta/\gamma \in [1.5, 8]$ (about 12% of prior mass removed; below 1.5 no epidemic occurs within the horizon, above 8 the peak precedes the first observation and β is weakly identified).
Inference is in $(\log \beta, \log \gamma)$.

Simulator $x_i = (x_{i,1}, \ldots, x_{i,10})$, $x_{i,t} \sim \mathrm{Binomial}\big(1000,\ \frac{I(t)}{N}\big)$
stored as fractions, where I is simulated from
$\frac{dS}{dt} = -\beta \frac{SI}{N}$
$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I$
$\frac{dR}{dt} = \gamma I$

Dimensions $\theta \in \mathbb{R}^2$, $x_i \in \mathbb{R}^{10}$

Fixed parameters	population $N = 10^6$, duration $T = 160$, initial conditions $(S(0), I(0), R(0)) = (N-1, 1, 0)$
Center	simulator method of moments: coarse grid, then three damped Gauss–Newton steps matching the simulator's mean output to the replicate mean
Reference posterior	grid integration in whitened Laplace coordinates; Gaussian to high accuracy, fitted contraction exponent 0.51
References	Kermack & McKendrick (1927)

C.8 Lotka–Volterra

The classic model of two interacting species; the task infers four interaction parameters from both populations observed at ten evenly spaced times.

Prior	$\alpha \sim \mathrm{LogNormal}(-0.125,\, 0.5)$, $\beta \sim \mathrm{LogNormal}(-3,\, 0.5)$, $\gamma \sim \mathrm{LogNormal}(-0.125,\, 0.5)$, $\delta \sim \mathrm{LogNormal}(-3,\, 0.5)$, truncated to parameters for which neither species falls below e^{-5} individuals at any observation time (about 8% of prior mass removed; an effectively extinct species carries no information about its rates). Inference is in log-rate coordinates.
Simulator	$x_{i,1,t} \sim \mathrm{LogNormal}(\log X(t),\, 0.1)$, $x_{i,2,t} \sim \mathrm{LogNormal}(\log Y(t),\, 0.1)$ stored as log-counts, where X, Y are simulated from $\frac{dX}{dt} = \alpha X - \beta XY$, $\frac{dY}{dt} = -\gamma Y + \delta XY$
Dimensions	$\theta \in \mathbb{R}^4$, $x_i \in \mathbb{R}^{20}$
Fixed parameters	duration $T = 20$ time units, initial conditions $(X(0), Y(0)) = (30, 1)$
Center	simulator method of moments: 13^4 grid, then four damped Gauss–Newton steps with multi-start
Reference posterior	adaptive importance sampling with a Student-t proposal at the Gauss–Newton MAP; Gaussian to high accuracy, fitted contraction exponent 0.50
References	Lotka (1920)

Simulator-based centers (SIR and Lotka–Volterra). Both centers are likelihood-free method-of-moments estimates computed per dataset from simulator runs only. *Objective:* the simulator's mean output (the ODE solution with observation noise off) is fitted to the replicate mean by weighted least squares, with weights n over the replicates' sample variance per output coordinate (floored at 10^{-3} of its mean), plus the Gaussian prior on θ as a quadratic penalty; the likelihood is never evaluated. *Initialization:* a coarse grid over the prior box, one batched ODE solve for the whole grid (41×41 points within ± 3.5 prior SDs for SIR; 13^4 points for Lotka–Volterra), scored by the same objective. *Refinement:* damped Gauss–Newton with a central-difference Jacobian and step halving, three steps from the best grid point for SIR and four steps from each of the eight best grid points for Lotka–Volterra, keeping the best fit (the oscillatory mean map has local optima at the wrong phase). The same routine is applied at every n; no out-of-range data, reference posterior or tuning enters the center.

The Gamma model is defined with the ablation that uses it (Appendix F).

D FULL REGULAR-BENCHMARK RESULTS

This appendix reports the six regular `sbibm` benchmarks at every evaluation size, in-range sizes next to the out-of-range sizes, under the unified protocol of Appendix A. The two symmetric multimodal benchmarks are in Appendix E.

Table 9: **Results on the six regular benchmarks at every evaluation size** (three seeds, except the mixture: five for C2ST and coverage, two for SD ratio; training $n \in \{5, \ldots, 20\}$). The two symmetric multimodal benchmarks are in Table 13. Bold marks the best entry in each column and block (closest to the target where the metric has one).

	C2ST				SD ratio				95% joint coverage			
estimator	$n=10$	$n=20$	$n=50$	$n=100$	$n=10$	$n=20$	$n=50$	$n=100$	$n=10$	$n=20$	$n=50$	$n=100$
Gaussian linear												
Free-G	0.57	0.61	0.75	0.89	1.02	1.09	1.24	1.41	**0.96**	0.97	0.97	0.86
Scale-G	0.55	0.57	0.65	0.71	1.01	**1.00**	**1.00**	**1.00**	**0.96**	**0.95**	0.91	0.87
CE-G	**0.51**	**0.51**	0.55	0.57	**1.00**	**1.00**	1.03	1.06	**0.96**	0.96	**0.96**	0.97
Free-NSF	0.55	0.58	0.72	0.87	1.01	1.06	1.26	1.53	**0.96**	0.98	0.99	1.00
CE-NSF	**0.51**	**0.51**	**0.54**	**0.56**	1.02	1.01	0.99	0.98	0.97	0.96	0.93	**0.93**
Gaussian linear (uniform)												
Free-G	0.84	0.88	0.94	0.98	1.62	1.97	2.72	3.62	**0.96**	0.98	0.90	0.55
Scale-G	0.66	0.67	0.68	0.74	1.06	1.03	1.02	**1.02**	**0.96**	**0.94**	0.92	0.90
CE-G	0.61	0.61	0.58	0.57	**1.03**	**1.01**	1.02	1.04	**0.96**	**0.94**	**0.95**	0.97
Free-NSF	**0.58**	0.62	0.78	0.92	**1.03**	1.12	1.35	1.62	**0.96**	0.98	0.99	0.99
CE-NSF	0.59	**0.59**	**0.56**	**0.56**	1.04	1.02	**1.00**	**0.98**	**0.96**	**0.94**	0.94	**0.94**
Gaussian mixture												
Free-G	0.87	0.90	0.93	0.95	2.73	3.34	4.54	5.86	0.93	0.93	0.97	0.98
Scale-G	0.87	0.89	0.90	0.92	2.81	2.83	2.74	2.70	**0.94**	0.89	0.81	0.75
CE-G	0.64	0.55	0.53	0.54	**1.22**	**1.03**	0.90	0.83	**0.94**	**0.94**	0.92	0.88
Free-NSF	0.87	0.89	0.93	0.95	3.14	3.44	4.08	4.76	**0.94**	**0.94**	0.93	0.86
CE-NSF	**0.56**	**0.52**	**0.51**	**0.51**	1.33	1.15	**1.05**	**1.01**	0.92	**0.94**	**0.94**	**0.93**
Bernoulli GLM												
Free-G	1.00	1.00	1.00	1.00	6.56	8.88	13.98	19.49	0.93	0.94	0.93	0.88
Scale-G	0.99	1.00	1.00	1.00	3.80	4.04	4.10	4.07	0.94	0.82	0.24	0.06
CE-G	0.70	0.67	0.66	0.68	**1.05**	**1.04**	1.07	1.10	0.92	0.94	**0.95**	0.97
Free-NSF	0.98	0.99	1.00	1.00	2.24	2.97	4.42	6.02	0.94	**0.95**	0.96	0.91
CE-NSF	**0.67**	**0.64**	**0.63**	**0.65**	**1.05**	**1.04**	**1.05**	**1.08**	**0.95**	**0.95**	**0.95**	**0.95**
SIR												
Free-G	0.88	0.92	0.97	0.99	2.27	3.07	4.37	7.43	0.94	**0.96**	0.61	0.36
Scale-G	0.88	0.90	0.96	0.99	2.30	2.60	3.09	4.46	**0.95**	0.92	0.46	0.15
CE-G	0.64	0.58	0.66	0.77	**1.13**	**1.08**	**1.27**	**2.07**	0.94	0.93	**0.95**	**0.97**
Free-NSF	0.74	0.79	0.92	0.96	1.46	1.80	2.44	3.78	**0.95**	0.97	0.63	0.38
CE-NSF	**0.60**	**0.56**	**0.61**	**0.74**	1.58	1.30	1.50	2.31	**0.95**	0.93	0.90	0.76
Lotka–Volterra												
Free-G	0.99	0.99	1.00	1.00	23.60	34.30	52.05	75.06	0.94	**0.96**	0.70	0.26
Scale-G	0.99	0.99	1.00	1.00	24.42	25.25	24.07	24.11	**0.95**	0.87	0.63	0.33
CE-G	0.74	0.71	0.67	**0.74**	**1.17**	**1.11**	**0.99**	0.90	0.94	0.93	0.89	**0.79**
Free-NSF	0.97	0.98	0.99	1.00	2.93	3.84	5.43	7.45	**0.95**	0.97	**0.90**	0.56
CE-NSF	**0.71**	**0.66**	**0.66**	0.76	1.57	1.43	1.19	**1.03**	0.97	**0.94**	0.84	0.66

E NONREGULAR STRESS TESTS

The four prototypes. Each prototype is a one- or two-dimensional model with an exact reference posterior, each chosen to probe one assumption of the structured representation:

- **T1, parameter-dependent support.** $x_i \sim U[0, \theta]$, $\theta \sim U[1, 3]$. The posterior is $\propto \theta^{-n}$ on $[\max(M_n, 1), 3]$ with $M_n = \max_i x_i$ and contracts at rate n^{-1} to a Gamma limit, the nonregular Bernstein–von Mises case of Bochkina & Green (2014). *Violated: the imposed rate* $1/2$.
- **T2, boundary parameter.** $x_i \sim \mathcal{N}(\theta, 1)$ with $\theta \in [0, 3]$ and a truncated-exponential prior of rate 2. The posterior is $\mathrm{TN}(\bar{x} - 2/n,\ 1/n,\ [0, 3])$: the rate $1/2$ holds, but the shape is non-Gaussian whenever the mass sits against the boundary at 0. *Probed: residual shape at a boundary, which the Gaussian head can represent after renormalization.*
- **T3, singular parameterization.** $x_i \sim \mathcal{N}(\theta_1\theta_2, 1)$, $\theta \sim U[-2, 2]^2$. The posterior concentrates on the hyperbola $\theta_1\theta_2 = \bar{x}$, a ridge that does not contract. *Violated: a single local center.*
- **T4, persistent bimodality.** $x_i \sim \mathcal{N}(\theta^2, 1)$, $\theta \sim U[-2, 2]$. The posterior has two modes at $\pm\sqrt{\bar{x}}$ whose separation does not shrink with n. *Violated: a single local center.*

Reference posteriors are computed on composite grids covering the prior support, and every estimator's density is renormalized on that support before scoring. The two symmetric `sbibm` benchmarks belong to the same family: SLCP has four mirror-image modes (as T4), and the support boundary of two moons contracts at rate n^{-1} (as T1).

Protocol. The protocol is that of the main text (Appendix A) with three differences: ten seeds instead of three, a smaller network (two hidden layers of 64 units, 3000 steps), and a nominal coverage level of 90% with C2ST evaluated in closed form on the reference grid. The structured estimators impose the same rate $\alpha = 1/2$ as on the regular benchmarks, and the center-estimated estimators use the closed-form center of each prototype (M_n for T1, $\max(\bar{x}, 0)$ for T2, the nearest point on the ridge for T3, the positive mode for T4); for T1 the encoder additionally receives M_n as a fixed summary, since mean pooling cannot represent a maximum. The true rate (1 on T1) is used only in explicitly labeled oracle ablations. One prototype-specific quantity is added to the main-text metrics: for T3 and T4, the distance from the posterior center to the nearest ridge or mode.

Table 10: **Nonregular prototypes at** $n = 100$ ($5\times$ the largest training size; ten seeds). Each block violates a different assumption of the structured representation.

	T1: parameter-dependent support				*T2: boundary parameter*			
method	KL	SD ratio	center err.	C2ST	KL	SD ratio	center err.	C2ST
Free-G	12.27	**1.09**	4.72	0.92	0.12	**1.00**	0.59	0.61
Scale-G	1.31	2.69	2.02	0.82	0.24	0.83	0.39	0.61
CE-G	1.11	2.71	**0.54**	0.78	**0.10**	0.81	**0.18**	**0.58**
Free-NSF	0.65	2.02	0.87	0.70	0.17	1.02	0.52	0.61
CE-NSF	**0.49**	2.82	1.81	**0.68**	0.18	0.84	0.24	0.59
	T3: singular parameterization				*T4: persistent bimodality*			
method	KL	SD ratio	center err.	C2ST	KL	SD ratio	center err.	C2ST
Free-G	2.34	0.84	**0.01**	0.90	2.28	0.85	0.01	0.86
Scale-G	7.65	0.36	**0.01**	0.95	5.01	0.39	**0.00**	0.93
CE-G	8.40	0.36	0.40	0.92	5.67	0.39	0.43	0.90
Free-NSF	**0.67**	**0.95**	0.11	**0.70**	**0.14**	**1.03**	0.01	**0.61**
CE-NSF	26.81	0.36	0.50	0.89	23.48	0.39	0.60	0.79

Rate violation: T1. T1 contracts at rate n^{-1} while every estimator imposes $n^{-1/2}$. The estimated posterior is therefore too wide, by a factor that grows with n, consistent with the rate sensitivity of Proposition 2. Table 11 then changes only the imposed exponent: the default $1/2$, the value 0.8 that the training-range estimate of Appendix F.4 supplies, and the oracle 1. Width, center error and KL

improve smoothly as the exponent approaches 1, while coverage stays close to nominal for the default and estimated exponents, because a too wide posterior fails gracefully, and falls slightly below it at the oracle exponent. An estimated rate therefore repairs most of the failure without being exact.

Table 11: Result on T1 at $n = 100$.

	CE-G					CE-NSF				
imposed α	SD ratio	center err.	cov. 90%	C2ST	KL	SD ratio	center err.	cov. 90%	C2ST	KL
$1/2$ (default)	2.71	0.54	0.97	0.78	1.10	2.82	1.81	0.97	0.68	0.49
0.8 (estimated)	1.40	0.25	**0.91**	0.72	0.67	1.42	0.59	**0.92**	0.59	0.14
1 (oracle ablation)	**0.92**	**0.15**	0.86	**0.69**	**0.48**	**0.92**	**0.23**	0.85	**0.56**	**0.03**

Boundary shape: T2. The fixed-θ_0 strata of Table 12 isolate the boundary; away from it ($\theta_0 = 0.5$) every estimator is accurate, and at it neither residual learns the truncated shape. This is a learning failure, not a representational one: every estimator's density is renormalized on $[0, 3]$ before scoring, and a Gaussian with mean $\bar{x} - 2/n$ and variance $1/n$ so renormalized equals the true posterior exactly. The exact posterior's own 90% HPD coverage of $\theta_0 = 0$ is 0.96 at $n = 20$ and 0.94 at $n = 100$, not the nominal 0.90, because the truth sits on the boundary.

Table 12: Fixed-θ_0 strata, $n = 20 \to 100$ (ten seeds). T2 at the boundary ($\theta_0 = 0$), with the exact posterior as reference; T3/T4, distance from each estimator's posterior mean to the nearest ridge or mode in units of its width, which grows with n for any single-center summary, the exact posterior mean included. Accuracy on T3 and T4 is read from Table 10.

	T2, $\theta_0 = 0$ stratum				T3	T4
estimator	SD ratio	center err.	cov. 90%	KL	ridge dist.	mode dist.
Exact posterior	1.00→1.00	0.00→0.00	0.96→0.94	0.00→0.00	–	–
Free-G	0.85→0.74	0.19→0.29	0.70→0.93	0.11→0.11	3.4→7.9	9.5→21.2
Scale-G	0.83→0.86	0.19→0.61	0.53→0.03	0.13→0.43	3.4→7.9	9.5→21.2
CE-G	0.83→0.79	0.19→0.22	0.60→0.58	0.11→0.16	3.4→6.3	9.2→12.7
Free-NSF	1.01→1.14	0.03→0.58	0.77→0.26	0.01→0.23	3.4→7.8	9.5→21.0
CE-NSF	0.97→0.90	0.08→0.18	0.68→0.68	0.02→0.07	3.3→5.3	7.7→8.1

Table 13: **The two symmetric multimodal `sbibm` benchmarks.**

	C2ST				MMD				95% joint coverage			
estimator	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
SLCP												
Free-G	0.98	0.98	0.99	0.99	0.064	0.073	0.084	0.093	0.96	0.99	**0.94**	**0.87**
Scale-G	0.98	0.98	1.00	1.00	0.058	0.142	0.363	0.532	0.97	0.92	0.20	0.00
CE-G	0.98	0.98	0.98	0.99	0.040	0.072	0.144	0.208	**0.95**	**0.94**	0.87	0.69
Free-NSF	**0.75**	**0.79**	**0.83**	**0.91**	**0.004**	**0.004**	**0.007**	**0.013**	0.92	0.98	**0.96**	0.82
CE-NSF	0.91	0.97	0.98	0.98	0.147	0.290	0.427	0.514	0.27	0.15	0.13	0.09
Two moons												
Free-G	0.98	0.98	0.99	0.99	0.129	0.141	0.165	0.197	**0.96**	0.99	**0.98**	**0.77**
Scale-G	0.98	0.98	0.99	0.99	0.109	0.223	0.475	0.660	0.97	0.97	0.81	0.09
CE-G	0.97	0.98	0.99	0.99	0.112	0.195	0.347	0.432	0.98	**0.95**	0.57	0.39
Free-NSF	**0.80**	**0.81**	**0.90**	**0.97**	**0.014**	0.012	**0.018**	**0.050**	**0.96**	0.99	0.89	0.41
CE-NSF	**0.80**	**0.81**	0.94	0.98	0.016	**0.011**	0.119	0.221	**0.96**	0.96	0.40	0.11

No single local center: T3, T4, SLCP and two moons. On T3 and T4 every single-center estimator degrades with n; the free spline estimator degrades least (C2ST 0.70 and 0.61 at $n = 100$ against 0.79–0.95 for the others, Table 10) but is not accurate either. On the two symmetric `sbibm` benchmarks (Table 13) every single-center estimator fails at every size, including inside the training range; the free spline estimator retains far more of the structure (MMD an order of magnitude lower) yet its C2ST at $n = 100$ is still 0.91 on SLCP and 0.97 on two moons, so it does not recover the posterior either.

F MECHANISM ABLATIONS AND MATCHED CONTROLS

Each subsection answers one question about the main-text comparison: whether the free baselines are under-trained, whether the ordering depends on the training maximum, whether the imposed rate matters once the center is held fixed, whether the exponent can be estimated from the training range, whether the residual network needs the size input, and how a compositional estimator compares.

F.1 BEST-EFFORT FREE BASELINES

The free baselines of the main text share the training budget of every other estimator. To check whether they lose because they are under-trained rather than because they cannot extrapolate, Free-G and Free-NSF were retrained on all six regular benchmarks under a best-effort protocol: the same architecture, prior-standardized parameters, the learning rate selected from $\{10^{-3}, 3 \times 10^{-4}, 10^{-4}\}$ by in-range validation NLL, three times the step budget and a longer early-stopping patience. No out-of-range data enter training, selection or calibration. Table 14 places the result beside the standard estimators at the training boundary and at $5\times$.

Within the training range, the extra effort narrows the gap on four benchmarks: on both Gaussian linear tasks, the Bernoulli GLM and SIR the best-effort baselines fit markedly better at the boundary, and on Gaussian linear with a uniform prior the best-effort Free-NSF matches the CE estimators there. On the Gaussian mixture and Lotka–Volterra, by contrast, the large in-range errors remain. Beyond the range the picture is the same on every benchmark: the best-effort baselines still widen and drift at $5\times$, their C2ST never falls below 0.83 while CE-NSF stays at or below 0.76, and on Gaussian linear the tuned Free-G loses most of its joint coverage. Extra training effort therefore narrows the in-range gap but does not change the out-of-range ordering.

F.2 A SECOND TRAINING MAXIMUM ON SIR

To check that the ordering of the main text is a statement about the relative size n/N rather than about the particular boundary $N = 20$, every estimator (and the two best-effort free baselines) was retrained on SIR with training sizes drawn uniformly from $\{20, \dots, 80\}$ and paired evaluation at $n \in \{40, 80, 200, 400\}$, with fresh reference posteriors at those sizes and three seeds. Table 15 reports both training maxima on the common n/N axis. This study reuses the CE arms of the development runs, whose center is the likelihood-based Fisher-scoring estimator, at both training maxima; their $N = 20$ rows therefore differ slightly from the main-text SIR rows, which use the likelihood-free moment center of Appendix C. The study was not rerun with the moment center because the ordering at $N = 20$ is the same under either center.

CE retains the best C2ST at $5N$ at both training maxima, but the absolute errors differ. At $N = 80$, CE-G has a width ratio of 1.16 and joint coverage of 0.86 at $5N$, compared with 1.83 and 0.96 at $N = 20$; CE-NSF undercovers at $5N$ in both settings. The best-effort free baselines improve in-range fitting but still deteriorate out of range. Increasing the training maximum therefore preserves the observed C2ST ordering without making either accuracy or coverage invariant to N.

F.3 DOES THE IMPOSED RATE MATTER AT A FIXED CENTER?

This study compares, on the Gaussian mixture, two fixed-center controls built on the same ten-step EM center, the full CE-G model, and a standardized, tuned Free-G baseline. The two controls keep the posterior mean exactly at the EM estimate (residual mean constrained to zero, no correction network) and differ only in whether the covariance carries the imposed $n^{-1/2}$ factor; CE-G is the full method; Free-G (standardized, tuned) is the free baseline with prior-standardized parameters and a learning rate selected by in-range validation. The study uses the main-text protocol (training sizes up to 20, evaluation at 20 and 100; Table 16).

With the center held fixed, imposing the contraction factor improves C2ST and KL at $5\times$ over letting the covariance scale freely, while joint coverage moves from 0.974 to 0.909 against a nominal 0.95; the rate parameterization therefore matters on its own for agreement with the reference, not for coverage. In this comparison the fixed-center, imposed-scale control has lower C2ST and KL at $5\times$ than the full CE-G model, so the study isolates a benefit of scaling at a fixed center but does not show

Table 14: **Best-effort free baselines beside the standard estimators** at the training boundary ($n = 20$) and at $5\times$ ($n = 100$); seed means over three seeds. Bold marks the best entry in each column and block (closest to the target where the metric has one).

	$n = 20$ (training boundary)				$n = 100$ ($5\times$)			
estimator	SD ratio	center err.	cov.	C2ST	SD ratio	center err.	cov.	C2ST
Gaussian linear								
Free-G	1.09	0.11	0.97	0.61	1.41	0.59	0.86	0.89
Free-G (best effort)	1.01	0.07	**0.96**	0.54	**0.98**	0.81	0.39	0.93
Free-NSF	1.06	0.11	0.98	0.58	1.53	0.53	1.00	0.87
Free-NSF (best effort)	1.03	0.08	**0.96**	0.54	1.42	0.44	1.00	0.83
CE-G	**1.00**	0.05	**0.96**	**0.51**	1.06	0.12	0.97	0.57
CE-NSF	1.01	**0.04**	**0.96**	**0.51**	**0.98**	**0.10**	**0.93**	**0.56**
Gaussian linear (uniform)								
Free-G	1.97	1.03	0.98	0.88	3.62	2.42	0.55	0.98
Free-G (best effort)	1.05	0.13	**0.96**	0.64	1.13	0.78	0.61	0.92
Free-NSF	1.12	0.15	0.98	0.62	1.62	0.74	0.99	0.92
Free-NSF (best effort)	1.04	0.09	**0.96**	**0.56**	1.37	0.48	0.98	0.83
CE-G	**1.01**	0.10	**0.94**	0.61	1.04	0.10	0.97	0.57
CE-NSF	1.02	**0.08**	**0.94**	0.59	**0.98**	**0.08**	**0.93**	**0.56**
Gaussian mixture								
Free-G	3.42	2.48	0.93	0.90	5.88	3.67	0.98	0.95
Free-G (best effort)	3.14	2.36	0.94	0.89	5.10	3.20	0.98	0.95
Free-NSF	3.76	2.48	0.94	0.89	5.41	4.60	0.86	0.95
Free-NSF (best effort)	3.75	2.34	**0.95**	0.89	6.82	3.68	0.98	0.96
CE-G	**1.14**	0.07	0.94	0.55	0.87	0.06	0.88	0.54
CE-NSF	1.20	**0.06**	0.94	**0.52**	**1.04**	**0.05**	**0.93**	**0.51**
Bernoulli GLM								
Free-G	8.88	7.00	0.94	1.00	19.49	16	0.88	1.00
Free-G (best effort)	2.84	1.95	**0.95**	0.99	5.84	4.82	0.77	1.00
Free-NSF	2.97	1.94	**0.95**	0.99	6.02	4.70	0.91	1.00
Free-NSF (best effort)	1.71	0.88	0.96	0.92	3.17	3.14	0.87	1.00
CE-G	**1.04**	**0.13**	0.94	0.67	1.10	0.12	0.97	0.68
CE-NSF	**1.04**	**0.13**	**0.95**	**0.64**	**1.08**	**0.08**	**0.95**	**0.65**
SIR								
Free-G	3.07	2.07	**0.96**	0.92	7.43	18	0.36	0.99
Free-G (best effort)	1.41	0.47	0.98	0.71	2.91	5.27	0.47	0.94
Free-NSF	1.80	0.70	0.97	0.79	3.78	8.07	0.38	0.96
Free-NSF (best effort)	1.32	0.37	0.98	0.68	2.71	4.80	0.44	0.94
CE-G	**1.08**	0.17	0.93	0.58	**2.07**	0.41	**0.97**	0.77
CE-NSF	1.30	**0.16**	0.93	**0.56**	2.31	**0.33**	0.76	**0.74**
Lotka–Volterra								
Free-G	34.30	22	**0.96**	0.99	75.06	47	0.26	1.00
Free-G (best effort)	29.73	20	0.97	0.99	65.84	45	**0.85**	1.00
Free-NSF	3.84	1.83	0.97	0.98	7.45	9.39	0.56	1.00
Free-NSF (best effort)	2.14	1.10	0.97	0.95	3.70	3.05	0.64	0.99
CE-G	**1.11**	**0.19**	0.93	0.71	0.90	**0.13**	0.79	**0.74**
CE-NSF	1.43	**0.19**	**0.94**	**0.66**	**1.03**	0.19	0.66	0.76

Table 15: **SIR at two training maxima** against the relative size n/N.

		C2ST				SD ratio				95% joint coverage			
N	estimator	$0.5N$	$1N$	$2.5N$	$5N$	$0.5N$	$1N$	$2.5N$	$5N$	$0.5N$	$1N$	$2.5N$	$5N$
20	Free-G	0.88	0.92	0.97	0.99	2.27	3.07	4.37	7.43	0.94	**0.96**	0.61	0.36
20	Scale-G	0.88	0.90	0.96	0.99	2.30	2.60	3.09	4.46	**0.95**	0.92	0.46	0.15
20	CE-G	0.57	0.55	0.64	0.76	**1.02**	**1.00**	**1.14**	**1.83**	0.94	0.92	**0.95**	**0.96**
20	Free-NSF	0.74	0.79	0.92	0.96	1.46	1.80	2.44	3.78	**0.95**	0.97	0.63	0.38
20	CE-NSF	**0.55**	**0.53**	**0.61**	**0.75**	1.47	1.26	1.77	2.40	0.94	0.93	0.93	0.70
80	Free-G	0.96	0.97	0.99	0.99	5.30	6.79	11.63	17.81	**0.95**	0.97	0.73	0.56
80	Free-G (best effort)	0.78	0.82	0.92	0.96	1.40	1.84	3.05	5.07	0.94	0.97	0.73	0.58
80	Scale-G	0.95	0.96	0.98	0.99	4.36	4.51	5.90	7.79	**0.95**	0.92	0.61	0.31
80	CE-G	0.58	0.58	**0.64**	**0.72**	**1.04**	**1.03**	**1.02**	**1.16**	0.94	**0.94**	**0.92**	**0.86**
80	Free-NSF	0.89	0.91	0.97	0.99	2.69	3.50	5.95	10.05	0.96	0.97	0.71	0.47
80	Free-NSF (best effort)	0.76	0.80	0.91	0.96	1.43	1.79	2.69	4.20	0.94	0.97	0.72	0.52
80	CE-NSF	**0.54**	**0.55**	**0.64**	**0.72**	1.60	1.47	1.94	2.59	0.94	**0.94**	**0.92**	0.74

an additional gain from the learned location adjustment: once the center is accurate and the rate is right, most of the gain is already obtained, which is the scope stated in Section 6.

Table 16: Fixed-center controls on the Gaussian mixture at the training boundary ($n = 20$) and $5\times$ ($n = 100$); means across three seeds.

method	C2ST at $n = 20$	KL at $n = 20$	C2ST at $n = 100$	KL at $n = 100$	cov. at $n = 100$
Free-G (standardized, tuned)	0.898	2.317	0.942	2.980	**0.966**
CE-G ($\log n$ residual)	**0.552**	**0.121**	0.537	0.096	0.870
Fixed EM center, free scale	0.574	0.174	0.623	0.328	0.974
Fixed EM center, imposed scale	0.558	0.137	**0.513**	**0.033**	0.909

The same two controls were run on SIR and Lotka–Volterra with the simulator-based moment center, under the main-text protocol and on the paired cases of Appendix D (Table 17, C2ST only). Imposing the contraction factor at a fixed center is again what matters: with the covariance free, C2ST rises with sample size on both tasks, whereas with the factor imposed it stays close to CE-G at every size and is slightly below it at $5\times$ (SIR 0.750 against 0.772, Lotka–Volterra 0.704 against 0.737). The learned location adjustment therefore adds nothing measurable on these two tasks either. The spline head does: CE-NSF has the lowest C2ST at every in-range size and on SIR at every size, but at $5\times$ on Lotka–Volterra it falls behind the fixed-center Gaussian control, and on both tasks its joint coverage at $n = 100$ is well below nominal (Appendix D), so the in-range benefit of residual flexibility does not carry over intact to the largest extrapolation.

Table 17: Fixed-center controls on SIR and Lotka–Volterra: C2ST at every evaluation size (training sizes 5 to 20); means across three seeds, lowest per column in bold.

	SIR				Lotka–Volterra			
estimator	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
Fixed center, free scale	0.653	0.642	0.789	0.897	0.768	0.806	0.886	0.940
Fixed center, imposed scale	0.634	0.572	0.647	0.750	0.729	0.704	0.656	**0.704**
CE-G	0.640	0.579	0.663	0.772	0.738	0.713	0.671	0.737
CE-NSF	**0.602**	**0.559**	**0.614**	**0.740**	**0.707**	**0.665**	**0.655**	0.757

F.4 Can the exponent be estimated from the training range?

CE-NPE imposes the contraction rate. When no asymptotic argument supplies it, the only data available at training time are simulations at the training sizes, and the natural estimate is the rate at which the statistical center itself contracts there: for each of $J = 64$ parameters drawn from the prior and each training size $n \in \{5, 8, 12, 16, 20\}$ we simulate $R = 32$ replicate datasets, compute the spread of c_{stat} across replicates, and fit $\log \text{SD} = a - \hat{\alpha} \log n$ by least squares. This needs no likelihood and no reference posterior. Table 18 reports $\hat{\alpha}$ on every benchmark next to the rate CE-NPE imposes ($1/2$ throughout) and the true asymptotic rate.

Three readings follow. On the regular simulators other than the mixture, the estimate is close to the regular rate. On T1 it gives 0.80, which is the finite-n contraction of the maximum over $n \in [5, 20]$ rather than the asymptotic 1; imposing 0.80 is exactly the estimated-rate row of Table 11, which repairs most of the T1 failure, so the estimate is useful where the asymptotic rate is unknown and wrong. On the Gaussian mixture, the estimate is 1.38, the narrow mixture component is identified between $n = 5$ and $n = 8$, and the center spread contracts much faster than $n^{-1/2}$ in this transient regime before settling to the regular rate. An estimator that trusted the training range would impose an exponent that over-contracts out of range. A training-range estimate is therefore a check on the model-class rate, not a substitute for it: when the two disagree, the disagreement itself indicates that the training range is not yet in the regime whose rate one wants to extrapolate.

Table 18: Contraction exponent estimated from the training range.

benchmark	imposed rate	true asymptotic rate	estimated α
Gaussian linear	$1/2$	$1/2$	0.50
Gaussian linear (uniform)	$1/2$	$1/2$	0.48
Gaussian mixture	$1/2$	$1/2$	1.38
Bernoulli GLM	$1/2$	$1/2$	0.46
SIR	$1/2$	$1/2$	0.56
Lotka–Volterra	$1/2$	$1/2$	0.48
T1 (support)	$1/2$	1	0.80
T2 (boundary)	$1/2$	$1/2$	0.43
T3 (singular)	$1/2$	ridge	0.44
T4 (bimodal)	$1/2$	$1/2$	0.47

F.5 Does the residual network still need n?

Gamma model and protocol. Training sizes are $\{20, 50, 100, 200, 500, 1000\}$ and evaluation sizes are $\{1000, 2000, 5000, 10000\}$. The encoder is a Deep Sets network over the log-transformed observations. The structured center is the log of the maximum-likelihood (equivalently method-of-moments) rate estimate plus a learned correction applied in rescaled units, $c_n = \log \hat{\theta}_{\mathrm{MLE}} + n^{-1/2}\Delta_\phi(s_n, \log n)$; the residual density on z_n is a one-dimensional rational-quadratic spline flow with linear tails (Durkan et al., 2019). Each diagnostic uses five seeds.

We hold the structured center and $\sqrt{n}$ transformation fixed and change only whether the residual network receives $\log n$, $1/n$, or no size input. As Table 19 shows, all three variants have a fitted exponent of 0.500 and nearly identical SD ratios, coverage, center error, and Kolmogorov–Smirnov (KS) statistics for simulation-based calibration (SBC) (Talts et al., 2018) at $10\times$. The $\log n$ context improves KL from 0.0023 without context to 0.0018; $1/n$ gives 0.0012. These differences concern finite-sample density fitting, while the contraction exponents remain nearly identical. Sample size is not unnecessary: n still enters the analytic transformation and the structured center. In this Gamma ablation, explicit neural context becomes secondary once the dominant contraction law is encoded.

Table 19: Matched residual-context comparison at $n = 10{,}000$ (means over five seeds; coverage is the nominal 90% region).

residual context	KL	SD ratio	90% cov.	center error	SBC KS	exponent
$\log n$	0.0018	0.991	0.886	0.061	0.039	0.500
$1/n$	0.0012	0.998	0.889	0.059	0.038	0.500
none	0.0023	0.992	0.888	0.065	0.041	0.500

F.6 A score-based compositional comparator on the six regular benchmarks

Compositional estimators handle a large n by learning the posterior score for a block of observations and summing block scores at test time, so every n is supported without any single estimator having

Table 20: **Score-based compositional estimator on the six regular benchmarks.** PF-NPSE (three seeds) with the published annealed Langevin sampler and with the GAUSS sampler of Linhart et al. (2026). Coverage requires a density and is not available for PF-NPSE.

	C2ST				SD ratio			
estimator	$n = 10$	$n = 20$	$n = 50$	$n = 100$	$n = 10$	$n = 20$	$n = 50$	$n = 100$
Gaussian linear								
Free-NSF	0.55	0.58	0.72	0.87	**1.01**	1.06	1.26	1.53
PF-NPSE (Langevin)	0.56	0.57	0.57	0.62	1.11	1.12	1.03	**0.98**
PF-NPSE (GAUSS)	**0.51**	0.52	0.55	0.59	**0.99**	**0.99**	**1.00**	**0.98**
CE-NSF	**0.51**	**0.51**	**0.54**	**0.56**	1.02	**1.01**	0.99	**0.98**
Gaussian linear (uniform)								
Free-NSF	0.58	0.62	0.78	0.92	**1.03**	1.12	1.35	1.62
PF-NPSE (Langevin)	0.60	0.61	0.64	0.72	1.10	1.14	1.06	1.04
PF-NPSE (GAUSS)	**0.57**	**0.59**	0.68	0.74	0.94	0.96	**1.00**	1.04
CE-NSF	0.59	**0.59**	**0.56**	**0.56**	1.04	**1.02**	**1.00**	**0.98**
Gaussian mixture								
Free-NSF	0.87	0.89	0.93	0.95	3.14	3.44	4.08	4.76
PF-NPSE (Langevin)	0.94	0.95	0.96	0.96	5.65	6.06	5.62	5.83
PF-NPSE (GAUSS)	0.93	0.94	0.95	0.96	4.24	4.27	5.19	6.10
CE-NSF	**0.56**	**0.52**	**0.51**	**0.51**	**1.33**	**1.15**	**1.05**	**1.01**
Bernoulli GLM								
Free-NSF	0.98	0.99	1.00	1.00	2.24	2.97	4.42	6.02
PF-NPSE (Langevin)	0.75	0.83	0.89	0.94	1.18	1.33	1.14	1.10
PF-NPSE (GAUSS)	0.71	0.79	0.88	0.93	**1.02**	1.15	1.09	**1.07**
CE-NSF	**0.67**	**0.64**	**0.63**	**0.65**	1.05	**1.04**	**1.05**	1.08
SIR								
Free-NSF	0.74	0.79	0.92	0.96	1.46	1.80	2.44	3.78
PF-NPSE (Langevin)	0.78	0.83	0.84	0.91	1.70	2.09	2.19	3.76
PF-NPSE (GAUSS)	0.65	0.68	0.75	0.81	**0.98**	**1.09**	**1.32**	**1.53**
CE-NSF	**0.60**	**0.56**	**0.61**	**0.74**	1.58	1.30	1.50	2.31
Lotka–Volterra								
Free-NSF	0.97	0.98	0.99	1.00	2.93	3.84	5.43	7.45
PF-NPSE (Langevin)	0.97	0.98	0.98	0.99	2.12	2.51	2.17	2.51
PF-NPSE (GAUSS)	0.94	0.95	0.97	0.98	**1.54**	1.73	1.85	2.12
CE-NSF	**0.71**	**0.66**	**0.66**	**0.76**	1.57	**1.43**	**1.19**	**1.03**

to extrapolate. The score network of PF-NPSE (Geffner et al., 2023) conditions on a block of up to 20 observations; at test time the n observations are partitioned into blocks of 20 plus a remainder, each block is scored by the same network given its block size, and the block scores are summed with the diffused prior score subtracted once per additional block. For conditionally independent observations, as on all six benchmarks, the factorization of the posterior into block posteriors is exact at noise level zero, but the sum of diffused block scores is an approximation at positive noise levels (Geffner et al., 2023; Linhart et al., 2026); at $n \leq 20$ the dataset forms a single block and no composition takes place, so only the two out-of-range sizes exercise it. We run the authors' public implementation under the main-text protocol: the same number of training datasets as one main-text estimator, each simulated with 20 replicates and randomly masked to a block size in $\{5, \ldots, 20\}$ (so about 1.6 times the replicates a main-text estimator sees), the analytic diffused prior of the authors' code, which ignores the mild prior truncation of SIR and Lotka–Volterra, in-range validation only, three seeds, and the first 200 of the paired evaluation cases at every size. Two samplers draw from the composed score: the published annealed Langevin sampler and the GAUSS sampler of Linhart et al. (2026), which corrects the composition with a Gaussian approximation. Neither provides a density, so coverage is not reported. Table 20 places both beside Free-NSF and CE-NSF.

Composition extrapolates well where the block posterior is easy to learn. On Gaussian linear, PF-NPSE keeps the correct width at every size and its C2ST at $n = 100$ is close to that of CE-NSF and far better than that of Free-NSF. On the uniform-prior variant and on SIR it lies between the free baseline and the center-estimated estimator. On the Gaussian mixture, Bernoulli GLM, and Lotka–Volterra it fails at every size, inside the training range as well as outside, as the free baselines do. The mechanism differs from that of free conditioning: on Bernoulli GLM the composed width stays within a third of the reference at every size while the center error, in posterior standard deviations, grows roughly as $\sqrt{n}$, so the composed posterior has approximately the right width around a center whose error does not contract. The GAUSS sampler improves width and C2ST over the published sampler on most benchmarks and sizes, most clearly on SIR, without changing the ordering. Among the estimators in Table 20, CE-NSF has the best C2ST at $n = 100$ on all six benchmarks, and PF-NPSE is the closest alternative only on Gaussian linear. Configuration and per-case results accompany the code.

G Real-data application: mutation rate and growth from 1000 Genomes

Population genetics supplies a real-data instance of the paper's setting in which the center is a classical moment construction rather than one of ours, and in which neither a ground-truth parameter nor a tractable reference posterior exists: in-model evaluation is truth-based, and real-data evaluation uses only reference-free criteria. The five main-text estimators are compared under the main-text protocol; nothing else changes.

G.1 Task and center

Under the neutral coalescent (Kingman, 1982) with exponential population growth, genomic window i yields the summary triple $x_i = (S_i, \pi_i, \eta_i)$ for k haplotypes: segregating sites, mean pairwise differences, and folded singletons. All three means are linear in the per-kb scaled mutation rate θ, while growth raises the rare-variant fraction η/S and lowers π/S, so a growth rate is identified jointly with θ from the moments. The growth parameter is the raw rate divided by four, which keeps the rescaled residual inside the spline head's core interval (Appendix A).

Prior	$\theta \sim U[0.2,\ 6]$ per kb; growth parameter $\sim U[0,\ 2]$ (raw rate $U[0, 8]$)
Simulator	coalescent tree of $k = 20$ under exponential growth (time-changed merging); mutations Poisson on branches, class counts ξ_b, then $S = \sum_b \xi_b$, $\pi = \sum_b b(k-b)\xi_b/\binom{k}{2}$, $\eta = \xi_1 + \xi_{k-1}$
Dimensions	$\theta \in \mathbb{R}^2$, $x_i \in \mathbb{R}^3$
Center	simulator method of moments in the Watterson tradition (Watterson, 1975): the growth rate from the two observed ratios $(\bar\pi/\bar S, \bar\eta/\bar S)$ against moment curves precomputed once by simulation, then θ from the scale $\bar S$; a grid lookup with quadratic refinement at query time, likelihood-free and contracting at $n^{-1/2}$
Reference posterior	none: the likelihood is intractable
References	Kingman (1982); Watterson (1975)

G.2 Real data

Summaries are computed from 1000 Genomes phase-3 chromosome 22 (The 1000 Genomes Project Consortium, 2015): $k = 20$ haplotypes (ten YRI individuals), biallelic PASS SNVs, non-overlapping 10 kb windows with strict-mask accessibility at least 0.9, giving 404 windows (150,223 segregating sites). An evaluation dataset of size n subsamples n windows at least 60 kb apart; 400 such overlapping datasets are drawn per size with a fixed seed shared by every estimator. The real ratios ($\bar\pi/\bar S = 0.258$ against the constant-size 0.282; $\bar\eta/\bar S = 0.355$ against 0.297) show the rare-variant excess of a growing population. Window independence is approximate and the neutral model is misspecified for real genomes; the criteria below do not condition on the model being correct.

G.3 Reference-free criteria

With no ground truth, each (estimator, n) is scored over the 400 subsampled datasets by:

- **Consistency**: the fraction of datasets whose two marginal central 95% intervals both contain the full-data moment fit ($\theta = 1.70$, raw growth 0.75, from all 404 windows), and the rms standardized distance to it (radius). Because the fit is estimated from windows that overlap the evaluation subsets and uses the same moment functional as the CE center, this is a consistency diagnostic that favours CE-NPE, not a test of nominal coverage.
- **Predictive**: the energy distance (Székely & Rizzo, 2013) between posterior-predictive window summaries (one simulated window per posterior draw, draws clipped to the prior box) and held-out real windows, on standardized summaries, over the first 64 datasets per size; this criterion is independent of the anchor.
- **Posterior readout**: the typical posterior of (θ, γ) at each n (Figure 4).

G.4 Results

In-model SBC and coverage give complementary diagnostics. Using simulation-based calibration (Talts et al., 2018), we assess posterior ranks alongside joint coverage (Table 22). At $5\times$ the largest training size the free heads are over-wide or rank-miscalibrated, Scale-G's learned center has substantial location error, and CE-G and CE-NSF have the lowest SBC KS statistics although their joint coverage is below nominal; Free-NSF has a radius and coverage closer to their targets but a larger SBC KS statistic. No single diagnostic therefore establishes posterior calibration (Section 4.2).

On the real genome CE-NPE maintains consistency with the full-data moment fit. The consistency of CE-G and CE-NSF with the full-data fit is flat from $n = 10$ to $n = 100$ while every other estimator's decays, their standardized distance to the fit stays constant while the others' grows, and their predictive energy distance matches the best free baseline within seed variability (Table 23). At $n = 100$ their posteriors sit on the full-data fit with a positive growth rate, as expected for an expanding population, whereas Scale-G halves the growth rate and shifts θ by more than a posterior SD (Figure 4).

Table 22: **In-model protocol, every size** (seed means): rms standardized truth distance (radius; target 1), joint 95% HPD coverage, and maximum marginal SBC KS; bold: within one seed SD of the best per column.

	$n = 10$			$n = 20$			$n = 50$			$n = 100$		
	rad.	cov.	KS	rad.	cov.	KS	rad.	cov.	KS	rad.	cov.	KS
Free-G	1.04	0.94	0.052	0.87	0.98	0.063	0.82	1.00	0.087	0.82	**0.98**	0.131
Free-NSF	0.98	**0.95**	**0.033**	0.92	0.98	0.062	0.92	0.97	0.110	**1.01**	**0.92**	0.224
Scale-G	**1.00**	**0.95**	0.042	1.06	0.96	0.077	1.34	0.87	0.167	1.64	0.75	0.181
CE-G	0.97	**0.95**	0.046	**0.98**	**0.96**	**0.053**	1.14	**0.94**	0.091	1.24	0.87	**0.075**
CE-NSF	**1.00**	**0.95**	0.046	0.95	**0.95**	**0.048**	**1.04**	0.91	**0.039**	**1.10**	0.86	**0.068**

Table 23: **1000 Genomes windows, every size** (seed means): consistency with and standardized distance to the full-data moment fit, and posterior-predictive energy distance to held-out windows; bold: within one seed SD of the best per column.

	$n = 10$			$n = 20$			$n = 50$			$n = 100$		
	cons.	rad.	energy	cons.	rad.	energy	cons.	rad.	energy	cons.	rad.	energy
Free-G	0.79	1.32	0.162	0.81	**1.32**	0.132	0.77	1.32	0.094	0.66	1.59	0.085
Free-NSF	0.80	1.28	0.158	0.79	1.52	**0.123**	0.72	1.63	0.090	0.58	1.92	**0.077**
Scale-G	**0.83**	1.21	0.158	0.77	1.44	0.129	0.58	1.87	0.095	0.28	2.79	0.085
CE-G	**0.83**	1.23	**0.153**	0.81	**1.32**	**0.122**	**0.85**	**1.22**	**0.087**	**0.83**	**1.23**	0.079
CE-NSF	**0.83**	**1.18**	0.160	**0.83**	**1.28**	**0.121**	**0.86**	**1.22**	**0.088**	**0.85**	1.28	0.082