WHAT A CHEAP SIMULATOR BUYS IN AMORTIZED INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Amortized simulation-based inference needs many simulations, so practitioners train on a cheaper approximate simulator and spend a limited budget of exact simulations where it seems most useful. Reported savings, not measured error, justify both choices. We measure the error, in stochastic gene-expression models whose exact posteriors are available. Three findings follow. First, mixing surrogate and exact simulations in one training objective shifts the estimator's target, with a posterior-mean bias that grows as $0.315\,(1-f)^{1.16}$ in the surrogate share $1-f$ ($r = 0.97$). At every budget we measure, down to 160 exact simulations, the mixture carries more bias and wider posteriors than training on the exact simulations alone and is never better calibrated, and a two-stage scheme that keeps the fidelities apart only ties the exact-only estimator. Second, placing the exact budget by burstiness is a weak lever. Even for a count surrogate whose error is concentrated, it improves on random placement by 0.003 to 0.012 in distance to nominal coverage, while reversing the same ordering costs 0.057. Third, a classifier between the two simulators, with no neural posterior estimator and no reference posterior, reproduces the ordering of the trained estimators over six surrogate-system pairs and separates missing support from wrong shape. Matched controls trace the corruption to a continuum standing in for a discrete low-copy state, and the account carries over to a three-state promoter, a stochastic epidemic and public single-cell data. In our systems the approximation often costs more than the exact solver it replaces.

1 INTRODUCTION

Simulation-based inference infers the parameters of a stochastic model when the likelihood is intractable but the model still generates samples (Cranmer et al., 2020). Neural posterior estimation (NPE) trains a conditional density estimator on pairs (θ, x) drawn from the prior and a forward simulator (Papamakarios & Murray, 2016; Greenberg et al., 2019). Amortization makes it both attractive and expensive, since one estimator serves every future observation but needs simulations covering the whole prior. When the exact simulator is costly, the standard move is to substitute an approximate one, and multifidelity methods make that substitution systematic (Prescott & Baker, 2020; Warne et al., 2022; Hikida et al., 2025; Krouglova et al., 2026). The object one can afford to simulate is then not the object that generated the data, and nothing in the training loss announces the mismatch. Misspecification degrades neural SBI without warning (Cannon et al., 2022; Hermans et al., 2022; Schmitt et al., 2023), but that literature studies a mismatch between simulator and world, not between two simulators one chooses to combine.

Our starting point is that this mismatch is not noise. NPE maximises the expected log density of θ given x under the distribution that generated its training pairs (Papamakarios & Murray, 2016), so a training set that mixes exact and surrogate simulations makes the estimator target the posterior of the mixture, a bias that does not shrink as the training set grows. Three questions follow. Does adding surrogate simulations to a fixed budget of exact ones help? Does it matter where the exact ones go? And can a cheap measurement tell a practitioner, before any estimator exists, which surrogate to trust and where? We answer the first two negatively for the naive uses of a surrogate and the third positively. Answering any of them needs a setting where the exact posterior is known, so that we measure the bias rather than infer it from savings. Stochastic gene expression provides one. The two-state transcription model (Peccoud & Ycart, 1995; Raj et al., 2006) has a closed-form stationary

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.

law and a tractable likelihood, and its textbook approximations (Gillespie, 2000; Schnoerr et al., 2017) give surrogates whose discrepancy is known. It is our instrument, not our subject. Our contributions are:

- **(C1) What a surrogate buys at equal exact budget.** We measure the mixture bias that Proposition 1 predicts, show that it scales with the surrogate share, and find no budget, down to 160 exact simulations, at which the mixture beats training on the exact simulations alone. The two-stage scheme of the multifidelity literature (Krouglova et al., 2026), which the proposition does not cover, ties the exact-only estimator (Section 4.2).
- **(C2) Placement is weaker than assumed.** Adaptive acquisition of exact simulations (Krouglova et al., 2026) assumes the surrogate's error is concentrated somewhere one can find. Where it is not concentrated, no placement scheme beats random, not even an oracle. Where it is, guided placement gains about as much as the seed spread, while reversing the ordering costs about five times the best gain (Section 4.3).
- **(C3) A pre-inference diagnostic that ranks surrogates.** A classifier between exact and surrogate simulations yields a discrepancy that reproduces the ordering of the trained estimators over six surrogate-system pairs and separates missing support from wrong shape, with no estimator and no reference posterior (Section 4.4).

The causal account that localises the corruption to a continuum standing in for a discrete low-copy state (Section 4.1) and the transfer to a second promoter, an epidemic and public single-cell data (Section 4.5) are instrument and scope rather than contribution. Our systems are small enough that exact simulation stays affordable, which is what lets us measure the damage; the target setting is the one where it is not.

2 Related work

Neural SBI and misspecification. Neural SBI learns a posterior or likelihood from simulated pairs (Cranmer et al., 2020; Papamakarios & Murray, 2016; Greenberg et al., 2019; Papamakarios et al., 2021; Tejero-Cantero et al., 2020; Lueckmann et al., 2021). A growing literature detects or robustifies it under simulator mismatch (Cannon et al., 2022; Ward et al., 2022; Schmitt et al., 2023; Huang et al., 2023; Kelly et al., 2024; Gao et al., 2023; Sun et al., 2026), checks calibration (Talts et al., 2018; Hermans et al., 2022), characterizes what a misspecified likelihood-free posterior converges to (Frazier et al., 2020), or corrects a trained estimator with a few labelled real observations (Wehenkel et al., 2025; Senouf et al., 2025; Ruhlmann et al., 2026); Kelly et al. (2025) review the strategies. These detectors need a trained estimator and do not say what the surrogate gets wrong, and the correction methods assume the exact process is observable but not simulable. We study the opposite case, where the exact simulator exists but is affordable only in limited quantity, which lets us measure the damage rather than correct it.

Multifidelity and surrogate-aware inference. Cheap simulations reduce the cost of expensive ones in multifidelity ABC for stochastic kinetics (Prescott & Baker, 2020; 2021; Warne et al., 2022), in multifidelity NPE, which pretrains on low-fidelity simulations, fine-tunes on high-fidelity ones and can acquire high-fidelity parameters adaptively (Krouglova et al., 2026), in multilevel neural SBI, whose correction terms remove the bias of the low-fidelity term with seed-coupled simulators (Hikida et al., 2025), and in surrogate-based amortized inference that propagates surrogate uncertainty (Scheurer et al., 2026). Krouglova et al. (2026) assume negligible low-fidelity cost and list optimal allocation under a fixed budget as open, and Hikida et al. (2025) derive a budget split from quantities they note are hard to compute. Both take for granted that the surrogate's error is unevenly distributed over the prior. Our statement about single-stage mixing shares the premise of Hikida et al. (2025); what we add is the measurement of the bias, of the two-stage alternative at equal exact budget, of what guided placement buys, and of the cost premise, which our systems invert.

Stochastic gene expression. The telegraph model has a Poisson-Beta stationary law (Peccoud & Ycart, 1995; Raj et al., 2006) and is routinely fit to single-cell data (Munsky et al., 2012; Larsson et al., 2019). Exact solutions of larger networks are costly (Munsky & Khammash, 2006), so inference relies on the chemical Langevin equation (CLE) (Gillespie, 2000) and the linear noise approximation

(LNA), whose accuracy deteriorates at low copy number (Grima et al., 2011; Schnoerr et al., 2017). Hybrid models with a discrete promoter are an established class (Jahnke, 2011; Hasenauer et al., 2014; Lin & Galla, 2016), moment-based inference misleads for bursty genes (Fröhlich et al., 2016; Munsky et al., 2018), count-family surrogates and neural inference already exist for bursting data (Cao & Grima, 2018; Carilli et al., 2024; Huang et al., 2024; Jiang et al., 2021), and snapshot identifiability of the rates is limited (Gu et al., 2025).

3 SETUP

Telegraph model. A single gene copy switches between an inactive and an active promoter state and transcribes only while active. With mRNA degradation rate 1 the parameters are $\theta = (k_{\text{on}}, k_{\text{off}}, k_{\text{syn}})$, and the stationary marginal of the chemical master equation (CME) is Poisson-Beta (Peccoud & Ycart, 1995; Raj et al., 2006),

$$x \sim \text{Beta}(k_{\text{on}}, k_{\text{off}}), \quad m \mid x \sim \text{Poisson}(k_{\text{syn}} x), \tag{1}$$

with $\mathbb{E}[m] = k_{\text{syn}} k_{\text{on}}/(k_{\text{on}} + k_{\text{off}})$ and $\text{Fano}[m] = 1 + k_{\text{syn}} k_{\text{off}}/[(k_{\text{on}} + k_{\text{off}})(k_{\text{on}} + k_{\text{off}} + 1)]$. Burstiness rises with low k_{on} and large burst size $k_{\text{syn}}/k_{\text{off}}$. Marginalizing (1) gives a tractable likelihood,

$$P(m{=}n \mid \theta) = \frac{k_{\text{syn}}^n}{n!} \frac{B(k_{\text{on}}{+}n,\, k_{\text{off}})}{B(k_{\text{on}}, k_{\text{off}})} \, {}_1F_1(k_{\text{on}}{+}n;\, k_{\text{on}}{+}k_{\text{off}}{+}n;\, -k_{\text{syn}}), \tag{2}$$

which we evaluate through Kummer's transform and use for reference posteriors. We sample the exact law from (1); finite state projection (FSP), the exact route a system without a closed form would need, matches (2) to a total-variation distance below 10^{-13} (Appendix B).

Three-state refractory promoter. To test a system without using its stationary distribution, the promoter moves along OFF $\rightleftharpoons$ POISED $\rightleftharpoons$ ACTIVE with generator Q and transcribes only in ACTIVE, with $\theta = (k_{01}, k_{10}, k_{12}, k_{21}, k_{\text{syn}})$. The count distribution has no convenient closed form, so the exact simulator is FSP on the augmented generator, which Gillespie SSA cross-checks. Its first two moments are closed-form for any promoter generator (Appendix D),

$$\mathbb{E}[m] = k_{\text{syn}} \pi_A, \qquad \text{Fano}[m] = 1 + k_{\text{syn}} \left(\left[(I - Q)^{-1} \right]_{AA} - \pi_A \right), \tag{3}$$

with π the stationary law of Q and A the transcribing state.

Surrogates and controls. The CLE relaxes the promoter to a continuous $g \in [0, 1]$ and treats promoter and mRNA as diffusions with Gaussian reaction noise, integrated by Euler-Maruyama with reflecting boundaries (Appendix B); in the three-state model it diffuses all three promoter occupancies. The LNA draws $m \sim \mathcal{N}(\mu, \sigma^2)$ with the exact μ, σ^2, rounded and clamped to be non-negative. Two matched controls isolate the mechanism. A negative binomial (NB) built from the *same* exact moments, with dispersion $\mu/(F{-}1)$ and success probability $1/F$, changes only the distributional family. A hybrid keeps the promoter switching discrete and applies the CLE only to the mRNA (Jahnke, 2011; Hasenauer et al., 2014). Unclamped, rounded and truncated LNA variants separate Gaussianity from non-negativity enforcement.

Inference and metrics. We place a log-uniform prior over rates (Appendix B) and reduce each 500-cell snapshot to six summaries (mean, variance, Fano factor, zero fraction, median, 90th percentile). We train one masked-autoregressive-flow NPE per simulator on 8,000 simulations with `sbi` (Tejero-Cantero et al., 2020) and evaluate all of them on one shared test set of 2,000 exact snapshots per seed, reporting means over five seeds unless stated otherwise. Binned by the empirical Fano factor of each snapshot in equal-count sextiles, we report 90% credible-interval coverage; overall we also report signed posterior-mean bias, SBC ranks (Talts et al., 2018; Hermans et al., 2022), posterior contraction (posterior over prior std), C2ST to reference posteriors obtained by MCMC on (2) (Lueckmann et al., 2021), and the posterior mass that leaks outside the prior box. We sample all estimators identically, directly from the flow.

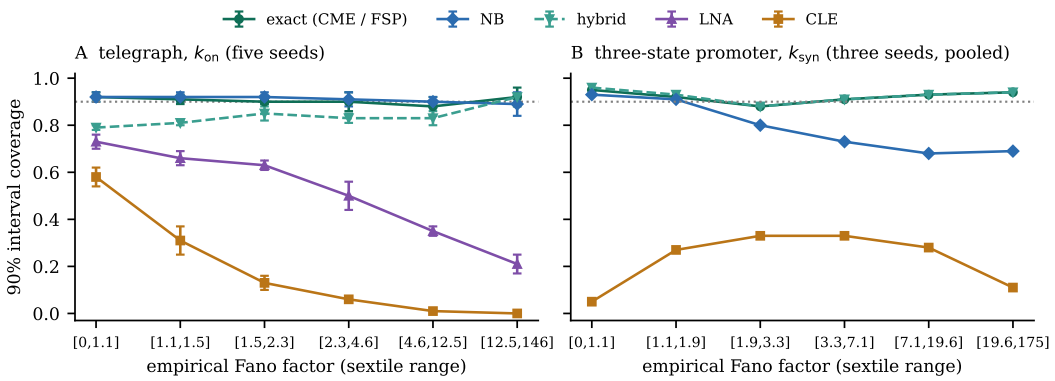


Figure 1: 90% interval coverage against the empirical Fano factor of each snapshot, in equal-count sextiles of 2,000 exact test snapshots, for NPEs trained on 8,000 simulations of each simulator. **A** Telegraph model, k_{on}, mean and one std over five seeds (Table 2); the hybrid ran on its own test sets. **B** Three-state promoter, k_{syn}, three seeds pooled (Table 3). The exact control stays at nominal (dotted), the hybrid close to it, the moment-matched LNA (A) and NB (B) degrade with burstiness, and the CLE fails throughout.

4 Results

4.1 Gaussian surrogates corrupt inference, and a discrete promoter repairs them

The instrument's error is directional. Sweeping k_{on} from 20 down to 0.025 at $k_{off} = 1$, $k_{syn} = 40$, the CLE overestimates the mean, suppresses the Fano factor and erases zero-inflation as burstiness grows; at the bursty end the exact zero fraction is 0.90 against 0.05 for the CLE, and the mean 0.97 against 5.7 (Figure 6, Appendix B).

These distortions reach the posterior (Figure 1A, Table 1). The CLE is miscalibrated in every sextile (k_{on} coverage 0.18 overall, 0.58 in the least bursty sextile and 0.00 in the burstiest). The LNA, built from the exact moments, is intermediate and degrades with burstiness, from 0.73 to 0.21, and its binned curve decreases in all five seeds (Appendix A). The CME control and the NB stay flat (0.88–0.92 and 0.89–0.92), so the decline belongs to the surrogate, not to the binning. Signed bias shows the direction, with the surrogates pushing k_{on} down (LNA -0.71, CLE -1.17 in $\log_{10}$ units) and k_{syn} up (LNA $+0.40$), and SBC ranks show the same one-sidedness (Appendix B). Posterior contraction shows that k_{off} is weakly identified for *every* estimator (0.77–0.87 of the prior width against 0.24–0.40 for the other rates), so similar k_{off} results reflect non-identifiability rather than recovery (Gu et al., 2025). C2ST to the reference posterior orders the estimators the same way above a floor of 0.70 that the control sets through the information the summaries discard, and the mass leaking outside the prior box grows with misspecification (0.13, 0.16, 0.39, 0.58 in the order of Table 1).

The CLE failure is promoter diffusion. The full CLE relaxes a single-copy promoter to a continuum. Keeping the promoter discrete and applying the CLE only to the mRNA repairs most of the damage: the hybrid covers k_{on} at 0.84 ± 0.01 (0.79–0.92 across sextiles) with signed bias -0.05, against 0.18 and -1.17 for the full CLE. On the sweep it cuts the fast-switching mean deficit of the CLE from -12.9% to -1.0% and tracks the exact zero fraction within 0.003 at every point, because the long promoter-off intervals that generate zeros survive. Its residual miscalibration does not track burstiness; it is worst in the least bursty sextile, which holds the fastest-switching promoters, where our one-switch-per-step integrator is least accurate (Appendix B).

The LNA failure is collapsing the promoter mixture into one Gaussian. The LNA has no promoter to diffuse, and continuity is not the problem, since the hybrid's mRNA is continuous too. A single Gaussian carrying the moments of a zero-inflated law must put mass below zero, so it either emits negative counts or distorts those moments. Both happen. With moments-only features an

Table 1: Prior-averaged inference quality on exact CME data (k_{on}; five seeds, 2,000 test snapshots per seed; seed std of the coverage at most 0.03). C2ST is to the reference posterior (three seeds). The hybrid ran on its own test sets, outside the scoring pipeline.

	CME	NB	LNA	CLE	Hybrid
coverage	0.91	0.91	0.51	0.18	0.84
signed bias	$+0.00$	-0.05	-0.71	-1.17	-0.05
C2ST	0.70	0.77	0.89	0.98	n/a
k_{off} contraction	0.77	0.87	0.80	0.78	n/a

unclamped Gaussian calibrates (0.92) while emitting 38% negative counts in a bursty example, and enforcing non-negativity breaks the moment match (clamping 0.45, truncation 0.32). With the full features the LNA stays at 0.51, whereas the NB, built from the same moments as a gamma-Poisson mixture, calibrates (0.91) because it represents the zero-inflated shape. The two controls restore discreteness where each surrogate lost it, explicitly in the hybrid and through the marginal family in the NB.

4.2 What a surrogate buys at equal exact budget

NPE maximises $\mathbb{E}[\log q_\phi(\theta \mid x)]$ under the distribution that generated its training pairs. If a fraction f of them come from the exact simulator, $x \sim p_{\mathrm{e}}(\cdot \mid \theta)$, and the rest from the surrogate, $x \sim p_{\mathrm{s}}(\cdot \mid \theta)$, that distribution is $p_{\mathrm{mix}} = f\, p_{\mathrm{e}} + (1-f)\, p_{\mathrm{s}}$.

Proposition 1. *Over conditional densities of unbounded capacity, the maximiser of the NPE objective on such a training set is the posterior under p_{mix}, $q^\star(\theta \mid x) \propto p(\theta)\, p_{\mathrm{mix}}(x \mid \theta)$. Unless $p_{\mathrm{s}} = p_{\mathrm{e}}$ almost everywhere, $q^\star$ differs from the exact posterior, and the difference does not depend on the number of training pairs.*

The proof is the usual argument that the expected log density is maximised by the conditional of the generating distribution. The statement is the premise of Hikida et al. (2025) and the reason Krouglova et al. (2026) keep the two fidelities in separate stages; we state it because it makes two testable predictions. The error should scale with the surrogate share and not shrink with more pairs, and it should trade off against the estimation error of a smaller exact budget.

The exact simulations alone beat the mixture. We simulate a fraction f of the 8,000 training parameters exactly and the rest with the LNA; all arms share seeds and test set. Since coverage above 0.90 is over-coverage, we compare distance to nominal coverage, signed bias and posterior width. Paired by seed, the mixture carries more k_{on} bias than training on the $f \cdot 8{,}000$ exact simulations alone by 0.182 ± 0.016, 0.117 ± 0.010 and 0.058 ± 0.016 at $f = 0.25$, 0.50 and 0.75 (five seeds), wider posteriors by 0.095, 0.060 and 0.045, and sits farther from nominal coverage by 0.012 ± 0.018, 0.004 ± 0.005 and 0.018 ± 0.013, a gap within the seed spread at the two smaller budgets (Figure 2B). The surrogate alone is not useless (0.505 ± 0.030 coverage), but it adds nothing to the exact simulations one can afford. The comparison fixes the *exact* budget, so it holds even if surrogate simulations were free.

The bias scales with the surrogate share, and there is no crossover. From $f = 0.02$ to 0.75 the mixture's k_{on} bias follows $|\mathrm{bias}| = 0.315\,(1-f)^{1.16}$ with $r = 0.97$ over seven budgets, consistent with proportionality to the surrogate share (Figure 2A). The predicted crossover, where a biased but plentiful training set should beat a small exact one, does not appear. At $f = 0.02$, that is 160 exact simulations, the exact-only estimator reaches coverage 0.910 ± 0.025 with bias -0.035 ± 0.015 and the mixture 0.703 ± 0.045 with bias -0.392 ± 0.138 and wider posteriors (0.574 against 0.350). Even 160 exact simulations already calibrate this task, so the regime where the bias would pay never materialises; we expect the crossover where the affordable exact budget is insufficient by itself, with more parameters or richer summaries.

Two stages avoid the damage without a gain. Pretraining on all 8,000 surrogate simulations and fine-tuning on the exact ones (Krouglova et al., 2026) never places both in one objective, so Proposition 1 does not apply. At equal exact budget it ties the exact-only estimator, with paired

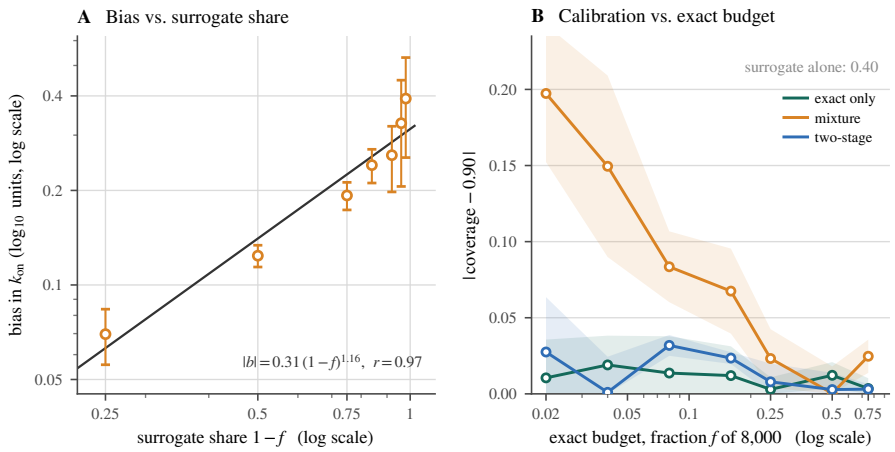


Figure 2: What a surrogate buys (telegraph model, LNA surrogate, 8,000 training simulations of which a fraction f is exact). **A** The mixture's k_{on} bias against the surrogate share over seven budgets ($f = 0.02$ to 0.75), with a power-law fit; the error scales with the share, the signature of bias that Proposition 1 predicts. **B** Distance to nominal k_{on} coverage against the exact budget. Training on the exact simulations alone is closest to nominal at every budget, the single-stage mixture is worse down to 160 exact simulations, and the two-stage scheme tracks the exact-only arm. Bands are one std over seeds; filled markers mark points with fewer than three seeds.

differences in distance to nominal of -0.001 ± 0.010, $+0.003 \pm 0.005$ and $+0.002 \pm 0.005$ at $f = 0.25$, 0.50 and 0.75, and within ± 0.005 in $|\mathrm{bias}|$ (fine-tuning rate in Appendix B).

4.3 Where to put the exact simulations

Given the exact budget, does it matter which parameters receive it? Five schemes decide, using the analytic Fano factor since θ is known at training time: *uniform* (random), *fano-hard* (the whole budget on the highest Fano), *fano-soft*(a) (sampling with weight $1 + a \cdot \mathrm{Fano\text{-}rank}$, $a \in \{2, 6\}$), and *inverse* (the lowest Fano, a control).

With the LNA, nothing beats random. No Fano-informed scheme improves on uniform beyond the seed spread (Table 5). The soft schemes match it, fano-hard is worse at low budget (0.76 against 0.88 at $f = 0.25$) and inverse is worse at every budget. An oracle that places exact simulations where the LNA-trained posterior mean is farthest from θ does no better (three seeds). Figure 1A says why: the LNA is miscalibrated in every sextile, so no region can be left to it.

With a concentrated error, the ordering carries signal but little margin. The NB is the favourable case. Its error is a shape failure confined to the bursty end (Section 4.4) and it damages k_{syn} (coverage 0.81), so we repeat the schemes with the NB and evaluate k_{syn}. At $f = 0.25$ the soft schemes improve on uniform by $+0.012 \pm 0.002$ and $+0.003 \pm 0.001$ in distance to nominal and fano-hard by $+0.007 \pm 0.003$ (two seeds, paired), while the reversed ordering costs -0.057 ± 0.004 (Figure 3). At $f = 0.50$ uniform already reaches 0.907 and every guided scheme overshoots (one seed). In the three-state promoter the NB repeats this scale (fano-soft(2) $+0.006$, inverse -0.071; one seed), and with the CLE there, concentrating the budget costs 0.259 while exact-only beats every mixed scheme by 0.123 (one seed). The Fano ordering carries real information, since reversing it hurts clearly, but placement converts it into gains comparable to run-to-run variability even in the favourable case.

4.4 Ranking surrogates before training any estimator

Both preceding sections judged surrogates from trained estimators, which is circular for a practitioner. A classifier removes the circularity. We simulate the exact and the surrogate simulator on the same prior draws and fit a classifier to tell them apart from (θ, x). Its AUC within a region estimates

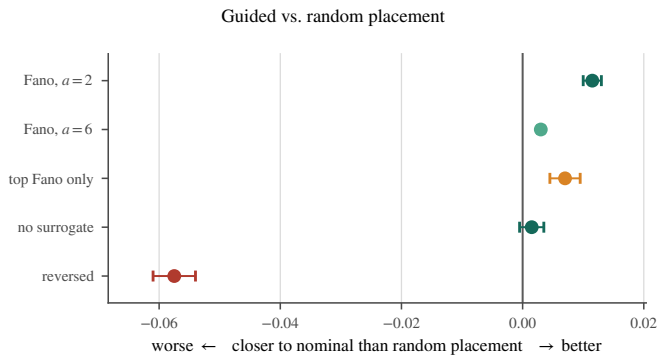


Figure 3: Placement with the concentrated NB surrogate (telegraph model, $f = 0.25$, two seeds). Each scheme chooses which quarter of the 8,000 training parameters gets the exact simulator; points give the change in distance to nominal k_{syn} coverage against random placement, paired by seed, positive meaning closer to nominal. Fano-weighted schemes gain little, while reversing the ordering loses about five times the best gain. *Fano*, a: weight $1 + a \cdot \text{Fano-rank}$; *top Fano only*: fano-hard; *no surrogate*: exact simulations alone; *reversed*: the ordering inverted.

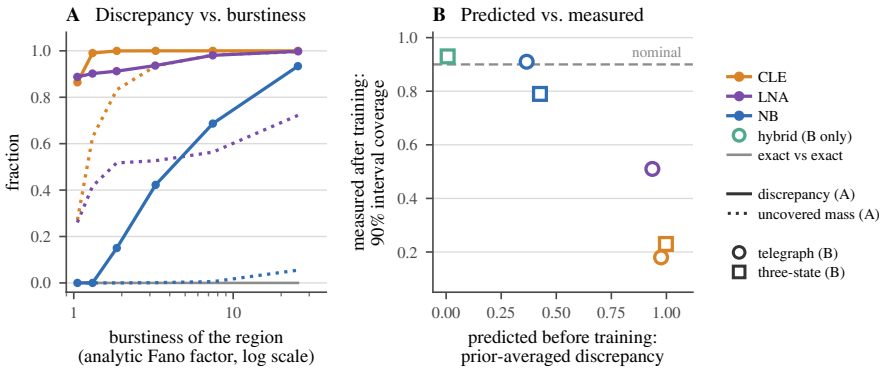


Figure 4: Ranking surrogates with no estimator. A classifier separates exact from surrogate simulations on shared prior draws (12,000 per simulator, three seeds). **A** Per burstiness bin of the telegraph model, the classifier discrepancy (solid) and the uncovered mass (dotted). The Gaussian surrogates are saturated everywhere and leave much of the exact mass unproduced, a support failure; the NB rises from zero with negligible uncovered mass, a shape failure confined to the bursty end. **B** Prior-averaged discrepancy against the 90% coverage each trained estimator attains (k_{on} in the telegraph model, k_{syn} in the three-state one); colour is the surrogate, shape the system. The ranking matches, with one inversion between the two saturated CLE entries.

the total-variation distance there, $\text{TV}_b \approx 2\,\text{AUC}_b - 1$, and the importance weights implied by its logit give the *uncovered mass*, the fraction of exact observations the surrogate essentially never produces. We bin by the analytic Fano factor and summarise each pair by the prior-averaged $\overline{\text{TV}}$ and the uncovered mass.

With 12,000 simulations per simulator and three seeds (Figure 4, Table 6), the ordering is hybrid (0.007), NB in the telegraph (0.366) and three-state models (0.426), LNA (0.936), CLE (0.976 and 0.998). It reproduces the ordering of the trained estimators' coverage, with one inversion between the two CLE entries, both at the ceiling. The exact simulator against itself gives 0.000 in every bin. The uncovered mass then separates two failure modes. The Gaussian surrogates fail on *support*, leaving 0.26–0.72 (LNA) and 0.27–1.00 (CLE) of the exact mass unproduced in every region, including the least bursty, so no region is safe. The NB fails on *shape*, with complete support (uncovered mass at most 0.055) and a discrepancy that grows from 0.000 to 0.934 with burstiness. This is what makes the NB the favourable case for placement, and it is available before any estimator exists. For the Gaussian surrogates TV saturates, so the uncovered mass carries the reading.

A per-observation statistic flags, but does not allocate. The snapshot Fano factor tracks the analytic one (correlation 0.94 ± 0.02 at 500 cells), so it orders observations by how far to distrust the LNA-trained posterior. Against a per-observation failure label (the LNA interval misses k_{on}), the empirical Fano factor reaches ROC AUC 0.67 ± 0.04 and the zero fraction 0.70 ± 0.06, both read from the data before inference, against 0.64 ± 0.04 for an out-of-distribution density detector (Schmitt et al., 2023) and 0.68 ± 0.05 for a posterior-predictive check. Paired by seed, both beat the density detector on average (the Fano factor in all five seeds), neither separates from the predictive check beyond the seed spread, and all are weak in absolute terms; flagging above a threshold fitted on seeds 0–2 marks 0.32 of observations and catches 0.45 of failures (Appendix A). Detection is not allocation, as Section 4.3 shows: a flag says where one trained estimator is untrustworthy, while allocation must make an amortized estimator trustworthy everywhere.

4.5 Transfer, real data and robustness

Three-state promoter. Against exact FSP test data (Figure 1B, three seeds), the exact-trained control is calibrated (k_{syn} coverage 0.92), the CLE collapses (0.23), and the hybrid tracks the control (0.93), more closely than in the telegraph model. The NB transfers only partially (0.79), degrading from 0.93 to 0.68 with burstiness, plausibly because the refractory marginal is further from a gamma-Poisson mixture. A count-appropriate family is therefore system-dependent, while the discrete-promoter hybrid is the robust surrogate in both systems. The CLE here is worst at both ends of the range (0.05 and 0.11), so a low Fano factor licenses trust in a shape-approximating surrogate, never in one that fails outright.

Stochastic epidemic. In an SIR model ($N = 200$, one initial infective), whether an outbreak takes off is again a discrete low-copy event. An NPE for R_0 trained on the exact SSA is calibrated (0.92), the CLE-trained one is not (0.68), and its failure sits in the middle of the Fano range rather than at the bursty end (one seed; Appendix E). The failure transfers; the ordering by burstiness does not.

Public single-cell data. On allele-resolved scRNA-seq of mouse embryonic stem cells (Larsson et al., 2019), with a binomial capture layer and 94 cells, the MMD between CME- and LNA-trained posteriors correlates with the observed Fano factor in both alleles (Spearman $\rho = 0.28$ on 6,430 C57 genes and 0.30 on 6,164 CAST genes, $p < 10^{-100}$), modestly at the median and clearly at the top, where the ten most-disagreeing genes per allele have Fano factors of 6.3 to 13.4 (Appendix F).

Robustness and cost. As the snapshot grows from 50 to 2,000 cells the control stays calibrated while the CLE worsens ($0.44 \to 0.13$), and from 2,000 to 128,000 training simulations the surrogates do not improve (LNA $0.65 \to 0.49$, CLE $0.18 \to 0.16$), as expected of a bias. Constraining the flow to the prior box removes the leakage without rescuing the surrogates (LNA 0.60, CLE 0.21), and a full count histogram with an embedding network preserves the ordering (CME 0.88, NB 0.85, LNA 0.61, CLE 0.22). The cost premise does not hold in our systems (Table 4). A training set costs 2.9 s from the closed form and 32 s by FSP against 6.8 min with the CLE in the telegraph model, and 64 s by FSP against 16.5 min in the three-state one; only the event-driven SSA is costlier.

5 Discussion and limitations

The practical summary is an order of operations. Estimate the discrepancy between the simulator pair first, since it is cheap, ranks surrogates correctly and says whether the failure is one of support or of shape. A support failure calls for replacing the surrogate, not scheduling around it. For a shape failure placement is defensible, but its margin over random placement is of the order of run-to-run variability, and concentrating the budget in the worst region is harmful. In either case, at equal exact budget and in these systems, the exact simulations one can afford are at least as good as any use of the surrogate, and mixing the two in one objective is worse. A surrogate for discrete low-copy systems should preserve the discrete switch, as the hybrid does, rather than match moments.

Limitations. (i) We use three to five parameters and six hand-built summaries; the crossover we do not find and the small placement margin may change in higher dimensions, where a surrogate's error may concentrate on smaller regions, which we regard as the main open question. (ii) The

placement experiment with the concentrated surrogate has two seeds in the telegraph model and one in the three-state promoter, so we report it as indicative. (iii) We set the diagnostic thresholds and the flagging rule after seeing these systems. (iv) Prior-averaged numbers depend on the prior box, which is why we stratify. (v) A tractable exact simulator is cheaper than or comparable to the CLE in all three systems, so we measure inference quality at a fixed exact budget rather than wall-clock savings; since Krouglova et al. (2026) assume negligible low-fidelity cost, this inversion is itself a finding about chemical-kinetics surrogates. (vi) On real data the effect is modest, the alleles are paired, the capture efficiency is assumed and the cells are few. We characterize rather than correct misspecification; correction with labelled real data is the complementary axis (Wehenkel et al., 2025; Ruhlmann et al., 2026).

6 Conclusion

Where the exact posterior is known, we can say what a cheap simulator is worth in amortized inference. Mixing it into the training objective moves the estimator's target, with a bias that scales with the surrogate share and that more training does not remove; keeping the fidelities in separate stages avoids the damage but produces no gain. Placing the exact budget by a statistic that genuinely ranks the surrogate's failures improves on random placement by little, although ordering it wrongly is expensive. What is cheap and worth doing first is the measurement itself, a classifier between the two simulators that ranks surrogates as the trained estimators do and says whether the failure is one of support or of shape. And in our systems, the time-stepped approximations cost more than the exact solver they replace.

Reproducibility Statement

The anonymized repository (`https://anonymous.4open.science/r/Genetic_SBI-8EFD`) regenerates every number in the paper from a clean clone, one script per figure and table, with the seeds, the cached simulation sets and the expected runtimes; seed counts are stated where each result appears. We unit-test the simulators against the closed-form moments and the ${}_1F_1$ zero fraction, and obtain reference posteriors by MCMC on the exact likelihood (2). Section 3 and Appendix B give all settings, and Appendix A the held-out seeds of the flagging rule.

Ethics Statement

This methodological study uses synthetic data from published models and a public single-cell dataset, involves no human subjects or private data, and has no foreseeable negative societal impact beyond the general risks of mis-applied statistical inference.

AI Use Statement

In this work we used generative AI tools to design and provide feedback on research methodology and experiments, to propose and refine hypotheses, to implement methods, to generate synthetic data sets, to clean and reformat the public data set we analyse, and to interpret results. We have not used generative AI tools to assist in the writing of proofs or to provide critical ingredients for proving mathematical claims. Among the tasks with recommended disclosure, we used generative AI tools to create and edit software code, to create and modify figures, to identify relevant literature and format references, to draft and edit parts of this paper, and to suggest its structure and title. We have reviewed all AI-assisted work: the simulators are unit-tested against closed forms, and the authors ran every experiment and checked every claim and number against the underlying results. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

Cannon, P., Ward, D. & Schmon, S.M. Investigating the impact of model misspecification in neural simulation-based inference. *arXiv:2209.01845*, 2022.

Cao, Z. & Grima, R. Linear mapping approximation of gene regulatory networks with stochastic dynamics. *Nat. Commun.* 9:3305, 2018.

Carilli, M. et al. Biophysical modeling with variational autoencoders for bimodal, single-cell RNA sequencing data. *Nat. Methods* 21:1466–1469, 2024.

Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *PNAS* 117(48):30055–30062, 2020.

Frazier, D.T., Robert, C.P. & Rousseau, J. Model misspecification in approximate Bayesian computation: consequences and diagnostics. *JRSS-B* 82(2):421–444, 2020.

Fröhlich, F. et al. Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.* 12(7):e1005030, 2016.

Gao, R., Deistler, M. & Macke, J.H. Generalized Bayesian inference for scientific simulators via amortized cost estimation. *NeurIPS*, 2023.

Gillespie, D.T. The chemical Langevin equation. *J. Chem. Phys.* 113(1):297–306, 2000.

Greenberg, D., Nonnenmacher, M. & Macke, J. Automatic posterior transformation for likelihood-free inference. *ICML*, pp. 2404–2414, 2019.

Grima, R., Thomas, P. & Straube, A.V. How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations? *J. Chem. Phys.* 135:084103, 2011.

Gu, J. et al. Scalable inference and identifiability of kinetic parameters for transcriptional bursting from single cell data. *Bioinformatics* 41(11):btaf581, 2025.

Hasenauer, J., Wolf, V., Kazeroonian, A. & Theis, F.J. Method of conditional moments (MCM) for the chemical master equation. *J. Math. Biol.* 69:687–735, 2014.

Hermans, J. et al. A trust crisis in simulation-based inference? Your posterior approximations can be unfaithful. *TMLR*, 2022.

Hikida, Y., Bharti, A., Jeffrey, N. & Briol, F.-X. Multilevel neural simulation-based inference. *NeurIPS 38*, 2025.

Huang, D. et al. Learning robust statistics for simulation-based inference under model misspecification. *NeurIPS*, 2023.

Huang, Z. et al. Deep learning linking mechanistic models to single-cell transcriptomics data reveals transcriptional bursting in response to DNA damage. *eLife* 100623, 2024.

Jahnke, T. On reduced models for the chemical master equation. *Multiscale Model. Simul.* 9(4):1646–1676, 2011.

Jiang, Q. et al. Neural network aided approximation and parameter inference of non-Markovian models of gene expression. *Nat. Commun.* 12:2618, 2021.

Kelly, R.P., Warne, D.J., Frazier, D.T., Nott, D.J., Gutmann, M.U. & Drovandi, C. Simulation-based Bayesian inference under model misspecification. *arXiv:2503.12315*, 2025.

Kelly, R.P. et al. Misspecification-robust sequential neural likelihood for simulation-based inference. *TMLR*, 2024.

Krouglova, A.N., Johnson, H.R., Confavreux, B., Deistler, M. & Gonçalves, P.J. Multifidelity simulation-based inference for computationally expensive simulators. *ICLR*, 2026.

Larsson, A.J.M. et al. Genomic encoding of transcriptional burst kinetics. *Nature* 565(7738):251–254, 2019.

Lin, Y.T. & Galla, T. Bursting noise in gene expression dynamics: linking microscopic and mesoscopic models. *J. R. Soc. Interface* 13(114):20150772, 2016.

Lueckmann, J.-M. et al. Benchmarking simulation-based inference. *AISTATS*, 2021.

Munsky, B. & Khammash, M. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* 124(4):044104, 2006.

Munsky, B., Neuert, G. & van Oudenaarden, A. Using gene expression noise to understand gene regulation. *Science* 336(6078):183–187, 2012.

Munsky, B. et al. Distribution shapes govern the discovery of predictive models for gene regulation. *PNAS* 115(29):7533–7538, 2018.

Papamakarios, G. & Murray, I. Fast ε-free inference of simulation models with Bayesian conditional density estimation. *NeurIPS 29*, 2016.

Papamakarios, G. et al. Normalizing flows for probabilistic modeling and inference. *JMLR* 22(57):1–64, 2021.

Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* 48(2):222–234, 1995.

Prescott, T.P. & Baker, R.E. Multifidelity approximate Bayesian computation. *SIAM/ASA J. Uncertain. Quantif.* 8(1):114–138, 2020.

Prescott, T.P. & Baker, R.E. Multifidelity approximate Bayesian computation with sequential Monte Carlo parameter sampling. *SIAM/ASA J. Uncertain. Quantif.* 9(2):788–817, 2021.

Raj, A. et al. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* 4(10):e309, 2006.

Ruhlmann, P.-L., Rodrigues, P.L.C., Arbel, M. & Forbes, F. Flow matching calibration for simulation-based inference under model misspecification. *ICML*, 2026.

Scheurer, S., Reiser, P., Brünnette, T., Nowak, W., Guthke, A. & Bürkner, P.-C. Uncertainty-aware surrogate-based amortized Bayesian inference for computationally expensive models. *TMLR*, 2026.

Schmitt, M. et al. Detecting model misspecification in amortized Bayesian inference with neural networks. *DAGM GCPR*, 2023.

Schnoerr, D., Sanguinetti, G. & Grima, R. Approximation and inference methods for stochastic biochemical kinetics: a tutorial review. *J. Phys. A* 50(9):093001, 2017.

Senouf, O., Wehenkel, A., Vincent-Cuaz, C., Abbe, E. & Frossard, P. Inductive domain transfer in misspecified simulation-based inference. *NeurIPS 38*, 2025.

Sun, S., Nicholls, G. & Lee, J. Amortized simulation-based inference in generalized Bayes via neural posterior estimation. *ICML*, 2026.

Talts, S. et al. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv:1804.06788*, 2018.

Tejero-Cantero, A. et al. sbi: a toolkit for simulation-based inference. *JOSS* 5(52):2505, 2020.

Ward, D. et al. Robust neural posterior estimation and statistical model criticism. *NeurIPS 35*, 2022.

Warne, D.J., Prescott, T.P., Baker, R.E. & Simpson, M.J. Multifidelity multilevel Monte Carlo to accelerate approximate Bayesian parameter inference for partially observed stochastic processes. *J. Comput. Phys.* 469:111543, 2022.

Wehenkel, A. et al. Addressing misspecification in simulation-based inference through data-driven calibration. *ICML*, 2025.

A MONOTONICITY OF THE COVERAGE CURVE AND THE PER-OBSERVATION FLAG

The ordinal claim the diagnostic uses is that high-Fano observations are less trustworthy than low-Fano ones. As supporting evidence, the binned LNA coverage curve decreases in every seed: the rank correlation between the Fano-sextile index and the per-sextile k_{on} coverage is -1.00 ± 0.00 over the five seeds for both the LNA and the CLE, with no ties in any seed. The two calibrated estimators show no such trend (CME $+0.09$ to -0.77, NB $+0.12$ to -0.58 across seeds), so the decline is not an artifact of the binning. This is a statement about the binned curve, not about individual test points, whose per-point coverage is a noisy $0/1$ indicator: adjacent bootstrap intervals can overlap even though the extreme sextiles (0.73 vs 0.21) do not. The diagnostic requires only the ordinal relation, not exact monotonicity.

Flagging rule. Because only the ordering matters, $\mathrm{Fano}^\star$ is an operating point rather than a universal threshold. We fix it as the Fano factor at which the smoothed LNA coverage crosses 0.45, fitted on seeds 0–2 (4.8; per seed 3.2 to 5.9); the same rule on the held-out seeds 3–4 gives 3.7. We chose the rule after seeing the fitting seeds, so the held-out seeds are what make it credible. Unflagged observations are not certified (the LNA sextiles below $\mathrm{Fano}^\star$ cover 0.50–0.73), so the flag sets priority for re-examination rather than licensing trust. At $\mathrm{Fano}^\star$ it marks 0.32 of observations and catches 0.45 of failures, where random flagging would catch 0.32.

Detector comparison. The failure label is that the LNA 90% interval misses the true k_{on} (failure rate 0.49 ± 0.05). By ROC AUC (five seeds), the empirical Fano factor reaches 0.67 ± 0.04, the zero fraction 0.70 ± 0.06 and the leaked posterior mass 0.70 ± 0.06, against 0.64 ± 0.04 for an out-of-distribution density detector (Schmitt et al., 2023) and 0.68 ± 0.05 for a posterior-predictive check under the surrogate. Since the seed spread is wider than the gaps, we also compare per seed on the same test points. The Fano factor beats the density detector by $+0.03 \pm 0.01$ in five seeds of five, the zero fraction by $+0.06 \pm 0.04$ in four and the leaked mass by $+0.06 \pm 0.03$ in five; against the predictive check the Fano factor draws (-0.005 ± 0.02) and the other two lead by $+0.02$. No statistic separates from the others beyond the seed spread (the zero fraction leads the Fano factor by $+0.03 \pm 0.04$), so burstiness carries the signal and the Fano factor is one way to read it. The snapshot statistics need no trained detector and no posterior, and the leaked mass vanishes for flows constrained to the prior box. The label is defined for the LNA; the CLE misses the truth even in the least bursty sextile (coverage 0.58), so a low Fano factor does not license trust in every surrogate.

B PER-BIN RESULTS, SIMULATORS, ARCHITECTURE, AND CONFIGURATIONS

Per-bin coverage. Table 2 gives the coverage curve of Figure 1A in numbers: per empirical-Fano sextile of the shared test set, the bin count and the k_{on} coverage of each estimator (mean $\pm$ std over five seeds); Fano ranges are sextile edges averaged over seeds.

Table 2: Per-sextile k_{on} 90% coverage (mean $\pm$ std over five seeds; n is the per-seed count, $2{,}000$ in total). Fano ranges are sextile edges averaged over seeds. The hybrid ran on its own test sets, whose sextile edges differ from these by under 3%.

bin	Fano range	n	CME	NB	LNA	CLE	Hybrid
0	$[0.0, 1.1]$	334	$0.92\pm.02$	$0.92\pm.02$	$0.73\pm.03$	$0.58\pm.04$	$0.79\pm.01$
1	$[1.1, 1.5]$	333	$0.91\pm.02$	$0.92\pm.02$	$0.66\pm.03$	$0.31\pm.06$	$0.81\pm.01$
2	$[1.5, 2.3]$	333	$0.90\pm.01$	$0.92\pm.02$	$0.63\pm.02$	$0.13\pm.03$	$0.85\pm.03$
3	$[2.3, 4.6]$	333	$0.90\pm.04$	$0.91\pm.03$	$0.50\pm.06$	$0.06\pm.01$	$0.83\pm.02$
4	$[4.6, 12.5]$	333	$0.88\pm.04$	$0.90\pm.02$	$0.35\pm.02$	$0.01\pm.00$	$0.83\pm.03$
5	$[12.5, 145.6]$	334	$0.92\pm.04$	$0.89\pm.05$	$0.21\pm.04$	$0.00\pm.00$	$0.92\pm.01$

Simulation-based calibration. Figure 5 gives the SBC rank histograms of one seed on $2{,}000$ exact test snapshots (in this seed, k_{on} coverage is 0.91, 0.92, 0.58 and 0.18 for CME, NB, LNA and CLE). The CME control is flat within the binomial 99% band. The NB is close to flat for k_{off}, tilts mildly toward the upper edge for k_{on}, consistent with its small negative bias, and piles up at the lower edge

Table 3: Three-state promoter: per-sextile k_{syn} 90% coverage, three seeds, $2{,}000$ exact FSP test snapshots per seed, pooled. The CLE is worst at both ends of the range; the moment-matched NB degrades with burstiness, as the LNA does in the telegraph model. Figure 1B plots these values.

bin	Fano range	n	exact	hybrid	NB	CLE
0	$[0.00, 1.13]$	334	0.95	0.96	0.93	0.05
1	$[1.13, 1.86]$	333	0.92	0.93	0.91	0.27
2	$[1.86, 3.26]$	333	0.88	0.88	0.80	0.33
3	$[3.26, 7.11]$	333	0.91	0.91	0.73	0.33
4	$[7.11, 19.61]$	333	0.93	0.93	0.68	0.28
5	$[19.61, 175.1]$	334	0.94	0.94	0.69	0.11
all		2000	0.92	0.93	0.79	0.23

for k_{syn}, the rate its coverage of 0.81 already shows it damages. The Gaussian surrogates deviate in the direction their signed bias predicts: ranks concentrate at the upper edge for k_{on}, meaning the posterior sits below the truth, and at the lower edge for k_{syn}, meaning it sits above. The CLE histogram for k_{syn} also has a second peak at the upper edge, the U-shape of posteriors that are too narrow, and for the weakly identified k_{off} the LNA is U-shaped while the CLE leans toward the upper edge. Taking the fraction of ranks in the outer tenth of the unit interval as a scalar summary (a calibrated estimator gives 0.10): for k_{on}, 0.09 (CME), 0.08 (NB), 0.42 (LNA) and 0.82 (CLE); for k_{syn}, 0.09, 0.18, 0.46 and 0.61; and for the weakly identified k_{off}, 0.08, 0.10, 0.27 and 0.23.

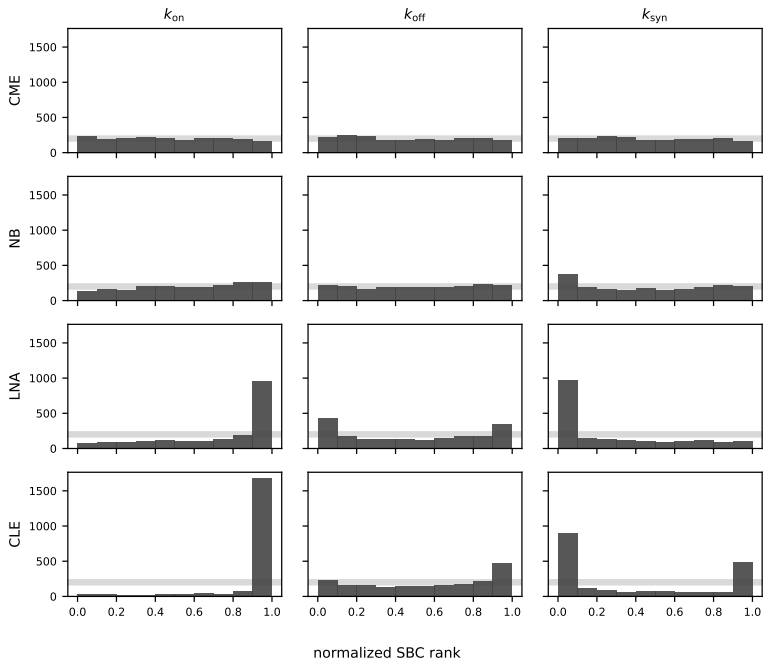


Figure 5: SBC rank histograms (rows: training simulator; columns: parameter), one seed and $2{,}000$ exact test snapshots; the grey band is the binomial 99% interval for a uniform histogram. The CME control is flat and the NB nearly so, apart from a peak at the lower edge for k_{syn}. The LNA and CLE pile up at the upper edge for k_{on} and the lower edge for k_{syn}, the one-sided signature of the signed bias in Table 1; the CLE is U-shaped for k_{syn}, the signature of posteriors that are also too narrow.

FSP truncation. We set the truncation per parameter from the analytic first two moments, $M = \lceil 60 + \mu + k_{\mathrm{syn}} + 15\sqrt{\sigma^2 + 1}\,\rceil$, and obtain the stationary distribution from the truncated generator. Over 200 parameters drawn from the prior it selects M between 76 and 1550 (median 114) and leaves at most 1.4×10^{-29} of the mass truncated; the resulting probabilities match (2) to a total-variation

distance of at most 7.2×10^{-14}, with relative errors of 6.5×10^{-14} in the mean and 6.5×10^{-13} in the Fano factor. For the three-state model the same rule is applied to the moments of (3).

Directional CLE error. Figure 6 gives the burst-frequency sweep of Section 4.1. At the bursty end ($k_{\text{on}} = 0.025$) the exact zero fraction of 0.90 matches the ${}_1F_1$ closed form to three decimals; at the fast-switching end ($k_{\text{on}} = 20$) the full CLE underestimates the mean by 12.9%, and the hybrid by 1.0%.

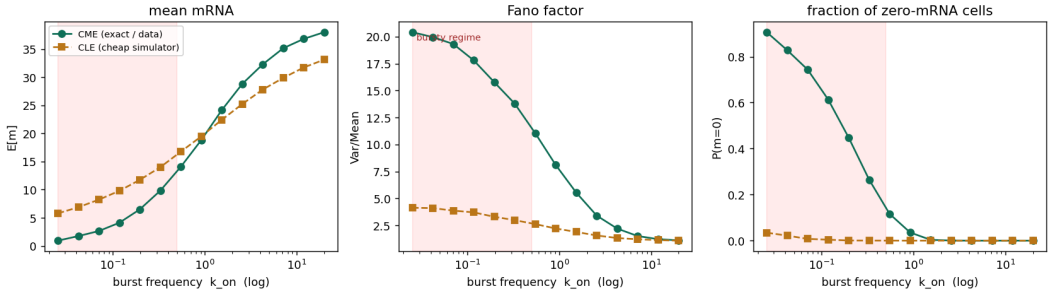


Figure 6: Directional CLE misspecification. Sweeping k_{on} log-uniformly from 0.025 to 20 at $k_{\text{off}} = 1$, $k_{\text{syn}} = 40$, the CLE (amber) is closest to the exact CME (green) when switching is fast, but as burstiness grows (shaded) it overestimates the mean, suppresses the Fano factor, and erases the zero-inflation (0.90 exact against 0.05).

Chemical Langevin equation. For the telegraph model the CLE reads

$$\begin{aligned} \mathrm{d}g &= (k_{\text{on}}(1-g) - k_{\text{off}}g)\,\mathrm{d}t + \sqrt{k_{\text{on}}(1-g)}\,\mathrm{d}W_1 - \sqrt{k_{\text{off}}g}\,\mathrm{d}W_2, \\ \mathrm{d}m &= (k_{\text{syn}}g - m)\,\mathrm{d}t + \sqrt{k_{\text{syn}}g}\,\mathrm{d}W_3 - \sqrt{m}\,\mathrm{d}W_4, \end{aligned} \tag{4}$$

with reflecting boundaries at $g \in \{0, 1\}$ and $m = 0$; in the three-state model the same scheme diffuses the three promoter occupancies on the simplex.

Simulator settings. We integrate the CLE and the hybrid by Euler-Maruyama with $\mathrm{d}t = 0.02$ up to $t = 20$ mRNA lifetimes; the CLE starts at the deterministic steady state, each cell is an independent trajectory, and the snapshot is the final state rounded to an integer. The hybrid keeps the promoter in $\{0, 1\}$, draws its initial state from the stationary law $\text{Bernoulli}(k_{\text{on}}/(k_{\text{on}} + k_{\text{off}}))$, and at each step flips it with the exact holding probability $1 - e^{-k\,\mathrm{d}t}$ of the two-state chain, so at most one switch occurs per step. At the fastest switching in the prior $k\,\mathrm{d}t \approx 0.4$, so the hybrid misses multiple switches within a step there; a $\mathrm{d}t$ sweep would settle whether this explains the hybrid's residual miscalibration (Section 4.1). The hybrid integrates the mRNA with the same Euler-Maruyama scheme as the CLE. For the NB, we raise Fano factors below $1 + 10^{-6}$ to $1 + 10^{-6}$, which makes it numerically Poisson. We set the Fano summary to 0 for an all-zero snapshot.

LNA variants (mechanism control, Section 4.1). The four LNA variants all draw a Gaussian with the exact mean and variance and differ only in the map to counts: *unclamped* (raw real values), *rounded* (nearest integer, may be negative), *truncated* (resample the negative tail), *clamped* (round then clip to ≥ 0, the LNA of the main text). For a gene with mean 2.86 and Fano factor 30 the unclamped and rounded variants (approximately, for rounding) preserve the exact moments but emit about 38% negative values, while truncation and clamping enforce non-negativity at the cost of inflating the mean to about 8.5 and 5.3 and compressing the variance. This is why the moments-only LNA failure is clamping, while the full-feature failure is the single-Gaussian marginal (Section 4.1).

Two-stage fine-tuning. The two-stage arm of Section 4.2 pretrains on all $8{,}000$ LNA simulations and fine-tunes on the exact ones. A fine-tuning rate of 3×10^{-5} under-corrects and leaves over-coverage (0.924), 3×10^{-4} slightly over-corrects (0.894), and we report 10^{-4}. Its paired differences to the exact-only arm in $|\text{bias}|$ are -0.003 ± 0.007, $+0.000 \pm 0.007$ and $+0.005 \pm 0.007$ at $f = 0.25$, 0.50 and 0.75. With the same seeds, the LNA-only estimator has k_{on} bias 0.632 ± 0.137 in magnitude, above the extrapolated intercept (0.315) of the power law, as expected since at $f \to 0$ the exact arm also degrades.

Table 4: Wall-clock per training set on one machine ($8{,}000$ simulations of 500 cells for the gene-expression models, $3{,}000$ of 200 outbreaks for SIR), extrapolated from 200 parameters (100 for FSP, 50 for SIR), all in one run. We time SSA per parameter, as median and interquartile range, because its cost scales with the number of reaction events and so spans a factor of 622 over the prior in the telegraph model and 923 in the three-state one, tracking the expected number of transcription events (log-log correlation 0.98 and 0.89). Absolute times depend on machine load; ratios within a run are stable to a few percent.

System	Simulator	kind	per training set
telegraph	Poisson-Beta	exact	2.9 s
telegraph	FSP (adaptive M)	exact	32 s
telegraph	SSA	exact	18 min (IQR 4–62)
telegraph	LNA / NB	surrogate	2.2 s / 2.7 s
telegraph	hybrid ($\mathrm{d}t = 0.02$)	surrogate	3.9 min
telegraph	CLE ($\mathrm{d}t = 0.02$)	surrogate	6.8 min
three-state	FSP (adaptive M)	exact	64 s
three-state	NB (moments by FSP)	surrogate	66 s
three-state	SSA	exact	2.2 h (median)
three-state	hybrid / CLE	surrogate	7.0 min / 16.5 min
SIR	SSA	exact	62 s
SIR	CLE	surrogate	57 s

Architecture and training. Each estimator is a masked autoregressive flow (5 transforms, 50 hidden features) trained through `sbi` with Adam at learning rate 5×10^{-4}, batch size 200, a 10% validation split, and early stopping after 20 epochs without validation improvement. Within each experiment all estimators share prior, features, architecture, and budget; only the training simulator differs.

Prior ranges. All telegraph-model estimators use a log-uniform prior: $\log_{10} k_{\mathrm{on}} \in [-2, 1.3]$, $\log_{10} k_{\mathrm{off}} \in [-1, 1.3]$, $\log_{10} k_{\mathrm{syn}} \in [0, 2.3]$. The three-state model uses $\log_{10} k_{01}, \log_{10} k_{12} \in [-2, 1.3]$, $\log_{10} k_{10}, \log_{10} k_{21} \in [-1, 1.3]$, and the same k_{syn} range. The SIR model has $N = 200$, $I_0 = 1$, $\gamma = 1$, and $\log_{10} R_0 \in [\log_{10} 0.5, \log_{10} 4]$.

Closed-form likelihood. We evaluate the Poisson-Beta marginal (2) stably in log space through Kummer's transform, ${}_1F_1(a; c; -\lambda) = e^{-\lambda}\,{}_1F_1(c - a; c; \lambda)$, whose positive argument makes the series a sum of positive terms; unit tests check it against a Monte Carlo estimate of the moments and against the ${}_1F_1$ zero fraction.

C Supplementary tables and figures

Table 5: Allocating exact simulation at equal budget: k_{on} coverage (mean $\pm$ std over seeds; nominal 0.90). Top: uniform, training on the exact simulations alone, and four Fano-informed schemes (five seeds). Bottom: oracle allocation by the surrogate's true error, with uniform rerun on the same three seeds. No allocation scheme is closer to nominal than uniform beyond the seed spread, and discarding the surrogate altogether is at least as good as the uniform mixture at every budget.

f	uniform	exact-only	fano-hard	fano-soft(2)	fano-soft(6)	inverse
0.25	$0.88 \pm .02$	$0.89 \pm .01$	$0.76 \pm .06$	$0.87 \pm .02$	$0.87 \pm .02$	$0.70 \pm .09$
0.50	$0.90 \pm .01$	$0.90 \pm .01$	$0.85 \pm .05$	$0.91 \pm .01$	$0.92 \pm .02$	$0.73 \pm .04$
0.75	$0.93 \pm .02$	$0.89 \pm .01$	$0.87 \pm .01$	$0.92 \pm .02$	$0.91 \pm .01$	$0.85 \pm .03$

	uniform	oracle-hard	oracle-soft
0.25	$0.87 \pm .01$	$0.84 \pm .00$	$0.88 \pm .02$
0.50	$0.90 \pm .01$	$0.91 \pm .03$	$0.92 \pm .01$
0.75	$0.92 \pm .01$	$0.92 \pm .01$	$0.92 \pm .01$

Table 6: Ranking surrogates before training any estimator (12,000 simulations per simulator, three seeds). $\overline{\mathrm{TV}}$ is the prior-averaged discrepancy from the classifier AUC; *gain* is the fraction of it that optimal placement could remove at half budget; *uncovered* is the mean fraction of exact observations the surrogate essentially never produces. The ordering of $\overline{\mathrm{TV}}$ reproduces the ordering of the trained estimators' calibration, except between the two saturated CLE entries, and the uncovered mass separates support failures from shape failures. Discrepancies are comparable within a system; across systems the trained estimators report different parameters (k_{on} and k_{syn}).

system	surrogate	$\overline{\mathrm{TV}}$	gain(0.50)	uncovered	failure
telegraph	exact (control)	0.000	–	0.000	none
telegraph	NB	0.366	0.86	0.010	shape
telegraph	LNA	0.936	0.04	0.501	support
telegraph	CLE	0.976	0.02	0.774	support
three-state	exact (control)	0.000	–	0.000	none
three-state	hybrid	0.007	–	0.000	none
three-state	NB	0.426	0.69	0.004	shape
three-state	CLE	0.998	0.00	0.928	support

D STATIONARY MOMENTS FOR ANY PROMOTER ARCHITECTURE

Let the promoter follow a continuous-time Markov chain with generator Q and stationary law π, transcribing at rate k_{syn} only in state A, with mRNA degradation rate 1. All reactions are zero- or first-order in m, so the moment equations close. Writing $a(t) = \mathbf{1}\{g(t) = A\}$, the stationary mRNA count is $m = \int_{-\infty}^{t} k_{\mathrm{syn}} a(s) e^{-(t-s)} \mathrm{d}s$ plus its birth-death noise, which gives

$$\mathbb{E}[m] = k_{\mathrm{syn}} \pi_A, \qquad \mathrm{Var}[m] = \mathbb{E}[m] + k_{\mathrm{syn}}^2 \int_0^\infty e^{-t} \,\mathrm{Cov}(a(0), a(t))\, \mathrm{d}t, \tag{5}$$

and since $\mathrm{Cov}(a(0), a(t)) = \pi_A (e^{Qt})_{AA} - \pi_A^2$ and $\int_0^\infty e^{-t} e^{Qt} \mathrm{d}t = (I - Q)^{-1}$, the integral evaluates to $\pi_A [(I - Q)^{-1}]_{AA} - \pi_A^2$, which is (3). Two checks. For the telegraph model, $Q = \left(\begin{smallmatrix} -k_{\mathrm{on}} & k_{\mathrm{on}} \\ k_{\mathrm{off}} & -k_{\mathrm{off}} \end{smallmatrix}\right)$ and (3) reduces exactly to the Fano factor of Section 3. For the three-state promoter, (3) agrees with the FSP moments to a relative difference of 9×10^{-13} over 200 parameter draws from the prior (FSP truncation $M = 600$, maximum truncated tail mass 3×10^{-120}). This is why the NB surrogate could be built in both systems at negligible cost (our three-state runs read its moments from the FSP solve instead, hence its cost in Table 4), and why its partial failure in the three-state model (Section 4.5) is about the shape of the marginal rather than about its moments.

E SIR PER-BIN COVERAGE

The SIR model has $N = 200$, $\gamma = 1$ and one initial infective; each observation collects the final sizes of 200 independent outbreaks, reduced to the same six summaries. The CLE integrates S and I by Euler-Maruyama with $\mathrm{d}t = 0.05$ and ends an outbreak when I reaches zero. We infer R_0 with NPEs trained on the exact SSA or on the CLE and evaluate them on SSA data in five equal-count bins of the empirical Fano factor (Table 7). The CLE is near-calibrated at the lowest and highest Fano and fails in between. The shape differs from the three-state model, where the CLE is worst at both ends; what they share is that a surrogate which fails outright is not ordered by burstiness. Without a hybrid with a discrete early phase we cannot say which part of the SIR approximation causes the failure.

F REAL-DATA GENES

On the allele-resolved scRNA-seq of Larsson et al. (2019) there is no ground-truth θ, so the checkable claims are that a CME-trained and an LNA-trained NPE disagree on some genes and that the observed Fano factor tracks where. The published CAST table lists 188 columns of which 94 are exact copies, so we use the 94 distinct cells matched to C57, keeping genes with mean count above 0.05 and a nonzero count in at least 20 cells; about 6% of entries are missing and we set them to zero, which inflates the zero fraction and hence the Fano factor. We train both estimators on 94-cell

Table 7: Stochastic SIR: R_0 90% coverage per equal-count bin of the empirical Fano factor of the final-size distribution (one seed, 1,000 test observations of 200 outbreaks each). The CLE fails in the middle of the range, not at its bursty end.

bin	Fano range	SSA (exact)	CLE
0	$[1.4, 10.4]$	0.94	0.85
1	$[10.4, 42.3]$	0.86	0.91
2	$[42.3, 61.4]$	0.89	0.29
3	$[61.4, 73.9]$	0.93	0.41
4	$[73.9, 90.1]$	0.97	0.95
all		0.92	0.68

snapshots with a capture layer, $\mathrm{obs} \sim \mathrm{Binomial}(\mathrm{true}, p)$ with an assumed $p = 0.2$, which maps $\mathrm{Fano}_{\mathrm{obs}} = 1 + p\,(\mathrm{Fano}_{\mathrm{true}} - 1)$; being monotone, it leaves the rank correlations unchanged. Per gene we compute the MMD between the two posteriors under a radial basis kernel with median-heuristic bandwidth on the pooled samples and correlate it with the observed Fano factor: Spearman $\rho = 0.28$ on C57 ($p = 9 \times 10^{-118}$, 6,430 genes) and $\rho = 0.30$ on CAST ($p = 2 \times 10^{-130}$, 6,164 genes). The effect is modest at the median and clearest at the top (Table 8), where *Zfp292* appears among the ten most-disagreeing genes of both alleles. Capture compresses the Fano factor monotonically, so it cannot explain why the effect is not larger, but its binomial noise over only 94 cells makes each gene's Fano estimate noisy.

Table 8: Ten most-disagreeing genes per allele, ordered by the MMD between the CME- and LNA-trained posteriors, with their observed Fano factor and mean count. All sit far above the median Fano of 2.2, and *Zfp292* is among the ten in both alleles.

C57	gene	Ndufb5	Rplp0	Prpf6	Traf7	Itga6	Fuca2	Zfp292	Rsrc1	Rpl5	Mrpl24
	Fano	9.6	8.7	7.8	7.4	6.3	7.4	13.4	6.7	8.8	6.6
	mean	5.2	3.7	4.1	4.1	2.6	3.0	8.2	2.4	3.7	3.2
CAST	gene	Tbc1d13	Lasp1	Impact	Sec62	Tcerg1	Ccdc88a	Fis1	Hnrnpf	Zfp292	Plod2
	Fano	9.2	7.4	7.9	12.2	8.0	8.3	7.6	6.6	11.5	6.3
	mean	4.6	4.2	4.5	6.5	3.6	3.0	3.0	3.2	5.4	2.8