EVALUATING MULTIMODAL INVERSE PROBLEMS: A PROTOCOL FOR PER-EVENT AND POPULATION-LEVEL FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluation in inverse problems is dominated by pointwise metrics – RMSE, MAE, per-event resolution – under the implicit assumption that lower error means better reconstruction. That the conditional mean poorly represents multimodal structure is well known but existing evaluation protocols do not score its consequences for the reported marginal. When one inverse problem is instantiated over many independent events and the scientific deliverable is the marginal $p(z)$ over the dataset, as in unfolding, a second failure mode appears that pointwise metrics cannot capture. By the law of total variance, point estimators trained to minimize MSE produce a marginal spectrum strictly narrower than the truth whenever the posterior has nonzero width. The resulting bias is independent of architecture, training, and dataset size, and it compresses precisely the spectral features like shapes, tails and modes that scientific measurements rely on. We propose a three-part evaluation protocol where each step targets a failure mode the others miss: per-event distributional accuracy via CRPS, population-level marginal accuracy via a spectrum-fidelity diagnostic, and uncertainty trustworthiness via coverage-based calibration. We instantiate the protocol on a synthetic benchmark with an analytic posterior to verify the mechanism, and on a realistic many-to-one inverse problem from particle physics we show that model rankings reverse between pointwise and distributional metrics, and calibration further separates architectures indistinguishable under CRPS. The choice of evaluation protocol can therefore determine the scientific conclusion.

1 INTRODUCTION AND SCOPE

When inferring underlying quantities from indirect or incomplete observations, evaluating the quality of the reconstruction is not a neutral postprocessing step: the metrics we report determine which models advance and shape the scientific conclusions drawn from them. In many such settings one inverse problem is instantiated over many independent events, and the reported result is an aggregate over them rather than any individual reconstruction. Nonetheless, evaluation is dominated by pointwise resolution metrics (RMSE, MAE, per-event bias) that ask how close each prediction lies to the truth. This convention fails structurally for under-constrained inverse problems with multimodal posteriors. The optimal predictor under mean-squared-error (MSE) is the conditional expectation (Bishop, 2006), which for multimodal posteriors can fall between modes in regions of vanishing probability density. Models achieving lower MSE do so by collapsing the posterior more aggressively, producing predictions that are individually "unphysical" and collectively distort reconstructed spectra. This is an *evaluation oversight* where the evaluation objective is misaligned with the scientific deliverable.

Scope We consider inverse problems where paired latents z and observations x are available at evaluation time, and the downstream quantity of interest is the marginal $p(z)$ over a dataset rather than a posterior over global parameters. We do not provide a full joint population-level evaluation for high-dimensional $z \in \mathbb{R}^d$; in Benchmark II, we complement the one-dimensional CRPS scores with the energy score, which evaluates predictive quality jointly across a full 8-dimensional latent space. Domain shift and unsupervised settings are outside scope.

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.

Table 1: Evaluation practices across target settings. Comparison of existing evaluation approaches and this work in terms of per-event distributional scoring, population-level marginal fidelity, and cross-family calibration.

	Per-event distributional score	Population marginal fidelity	Cross-family calibration
Aggregated-latent setting: per-event z, deliverable is $p(z)$			
Unfolding	n/a[a]	Yes	No
This work	**Yes**	**Yes**	**Yes**
Other settings: single observation, or posterior over global parameters			
Decision theory / proper scoring rules	Yes[b]	No	Partial
SBI coverage diagnostics	No	n/a[c]	Yes
SBI benchmarks; InverseBench	Partial	No	Partial
Distributional eval. in inverse problems	Yes	No	Partial

[a] Produces no per-event estimate. [b] CRPS, energy score. [c] Targets global parameters.

Within this setting, we present the following contributions:

1. A theoretical argument (section 3.2) that conditional mean collapse, itself standard, systematically mismodels aggregate spectra: any point estimator minimizing MSE produces a marginal spectrum strictly narrower than the truth whenever the posterior has nonzero variance – a bias independent of architecture, training objective, and dataset size.
2. A three-part evaluation protocol (per-event distributional accuracy, population-level spectrum fidelity, and coverage-based calibration) that diagnoses the failure modes pointwise metrics miss and applies across regression, mixture, and generative model families on a common scale: axes I and II without qualification, axis III among models supplying per-event uncertainty. We show the three axes are non-redundant: each axis detects a failure mode the others miss.
3. A verification of the protocol on a controlled synthetic benchmark with an analytic posterior and an empirical demonstration on a realistic many-to-one inverse problem from particle physics, showing that model rankings reverse between pointwise and distributional metrics, and that calibration further separates architectures indistinguishable under the per-event distributional accuracy metric.

2 Background and Related Work

Fully-Supervised Inverse Problems We consider the task of recovering a latent state $z \in \mathbb{R}^n$ from noisy observations $x \in \mathbb{R}^m$ related by a forward model $x = G(z, \xi)$, where ξ represents measurement noise or physical stochasticity – such a pair, (z, x), is commonly called an *event* or a *sample*. In scientific domains – particle reconstruction, seismic imaging, medical tomography – G is a high-fidelity simulator that is analytically intractable to invert. When a large corpus of paired (z, x) samples is available, the task becomes a *fully-supervised inverse problem*: learn a surrogate $G^\dagger$ approximating the posterior $p(z \mid x)$. Such problems are frequently ill-posed, admitting multiple or no exact solutions, either through the stochasticity of ξ or through information loss in x.

Evaluation in Scientific Reconstruction Because reconstruction is performed event-by-event, evaluation has traditionally followed the same structure: RMSE and MAE are the headline metrics in particle reconstruction (Ehrke et al., 2023; Shmakov et al., 2022), sparse-view computed tomography (Sidky & Pan, 2022), image super-resolution (Wang et al., 2021), climate downscaling (Vandal et al., 2018), and geophysical inversion (Kim & Nakata, 2018), and are routinely adopted as the training objective itself. Lower error is implicitly assumed to mean better reconstruction, also when aggregated. This holds for well-constrained unimodal problems, but not when the posterior has nonzero width (section 3). Point summaries are reported even when a full posterior is available, e.g. for gravitational-wave parameter estimation (Shen et al., 2021; Dax et al., 2021) and neutrino kinematics (Raine et al., 2024).

Connection to Simulation-Based Inference (SBI) Learning $G^\dagger$ is an instance of amortized neural posterior estimation (NPE) (Cranmer et al., 2020; Papamakarios & Murray, 2016). Our setting differs from general SBI in two respects: the latent z is per-event and amortized over $\mathcal{O}(10^4\text{-}10^6)$ events, and the scientific quantity of interest is the aggregated marginal $p(z)$ rather than a posterior over global parameters.

Unfolding in Physics Correcting a measured spectrum for detector effects is a general aggregated-latent problem and modern generative unfolding (Huetsch et al., 2025; Bellagente et al., 2020; Petitjean et al., 2025) learns this correction with the same model families we evaluate in this work. Its target is an ensemble, so evaluation asks whether that ensemble matches the truth. Our setting instead scores a per-event reconstructor, and requires that its aggregate match the marginal as well.

Proper Scoring Rules, Calibration, and Evaluation Principles A scoring rule $S(F, z)$ is *proper* if its expectation under the true distribution is minimized by F equal to that distribution, and *strictly proper* if uniquely so (Gneiting & Raftery, 2007). Propriety prevents metrics from being gamed by posterior collapse. The continuous ranked probability score (CRPS) is the standard strictly proper score for univariate prediction (Gneiting & Raftery, 2007; Hersbach, 2000):

$$\mathrm{CRPS}(F, z) = \int_{-\infty}^{\infty} \left[F(t) - \mathbf{1}(t \geq z)\right]^2 \mathrm{d}t, \tag{1}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. When the predictive distribution is represented by N posterior samples CRPS can be estimated in $\mathcal{O}(N \log N)$ by sorting (Jordan et al., 2019) (appendix A.1). Calibration assesses whether nominal credible regions contain the truth at their stated frequencies (Guo et al., 2017; Kuleshov et al., 2018). SBI benchmarks and coverage diagnostics (Lueckmann et al., 2021; Hermans et al., 2022; Talts et al., 2020; Lemos et al., 2023) develop these tools for global parameters, and conformal variants (Gammerman et al., 1998; Wang et al., 2023) have been applied in high-energy physics (Araz & Spannowsky, 2025). InverseBench (Zheng et al., 2025) establishes strong reconstruction baselines using primarily pointwise quality metrics and notes that assessing multimodal posterior structure remains open. Table 1 summarizes how these lines of work divide by target setting.

3 Limitations of Pointwise Evaluation Metrics

Pointwise metrics are misaligned with the goals of multimodal reconstruction. The arguments below are model-independent, applying to any method that extracts point predictions from a multimodal posterior. A complementary Jensen-gap failure for nonlinear observables is treated in appendix D.8.1.

3.1 The Conditional Mean Pathology

For a generic inverse problem with posterior $p(z \mid x)$, the point estimator minimizing the MSE is the conditional mean $\mathbb{E}[z \mid x]$ (Bishop, 2006). This is well known from decision theory, but becomes problematic when $p(z \mid x)$ is multimodal where the conditional mean may not correspond to a plausible posterior sample. Consider for example the symmetric bimodal posterior

$$p(z \mid x) = \tfrac{1}{2}\mathcal{N}(z;\, +a,\, \sigma^2) + \tfrac{1}{2}\mathcal{N}(z;\, -a,\, \sigma^2), \qquad a \gg \sigma. \tag{2}$$

The MSE-optimal prediction is $f_\theta(x) = 0$, which has posterior density $p(z = 0 \mid x) \approx 0$ whenever the modes are well separated ($a \gg \sigma$). In other words, the "best" prediction under the squared error is a value that the posterior assigns essentially zero probability.

3.2 Spectral Mismodeling under Aggregation

This per-event pathology has population-level consequences: in many applications the quantity of interest is the marginal distribution of z across the dataset, a *spectrum* or an *unfolded distribution*.

By the law of total variance (Grimmett & Stirzaker, 2020):

$$\mathrm{Var}[z] \;=\; \underbrace{\mathbb{E}\left[\mathrm{Var}\left[z \mid x\right]\right]}_{\text{within-posterior variance}} \;+\; \underbrace{\mathrm{Var}\left[\mathbb{E}\left[z \mid x\right]\right]}_{\text{variance of point estimates}}, \tag{3}$$

from which it follows immediately that

$$\text{Var}\left[\mathbb{E}[z \mid x]\right] \;\leq\; \text{Var}[z], \tag{4}$$

with equality only when $\text{Var}[z \mid x] = 0$ almost surely, i.e. when the true spectrum has no within-posterior variance. It states that the distribution of conditional means is *always narrower* than the true marginal distribution of z whenever $\text{Var}[z \mid x] > 0$.

The compression in eq. (4) is a *bias, not a variance term*: it does not decrease with dataset size and is inherited by any estimator converging to $\mathbb{E}[z \mid x]$. It has *direct physical consequences*: tails and shapes of marginal distributions encode physical parameters, so their compression directly biases downstream measurements. Population-level regularizers such as MMD penalties added to an MSE objective reduce but cannot eliminate the bias (see section 6): they reshape the aggregate of point predictions without correcting their per-event placement, leaving the within-posterior variance term of eq. (3) unmodelled. Finally, the compression is largest when multimodality is pronounced, since the within-posterior variance includes the spread between modes. MAE-trained models target the conditional median, the variance bound does not generally hold, but the median remains a point estimate that discards the within-posterior spread.

4 Proposed Three-Part Evaluation Protocol

We propose a three-metric protocol – per-event distributional accuracy (CRPS), population-level spectral accuracy (spectrum fidelity), and uncertainty calibration (coverage) – where each metric targets a failure mode the others miss and no two are interchangeable (section 4.4). Axes I and II apply to every model family; axis III requires per-event uncertainty. See appendix A for more details.

The axes score a distribution on a scalar quantity, either the latent z itself or a derived observable $\mathcal{O} : \mathbb{R}^d \to \mathbb{R}$. The metrics are unchanged in either case, and only the scored quantity differs. The choice is set by downstream use, since matching $p\left(\mathcal{O}\left(z\right)\right)$ does not imply matching $p(z)$.

4.1 Metric I: CRPS (Per-Event Distributional Accuracy)

Our per-event metric is the CRPS (section 2). Three properties make it the right choice for evaluating multimodal inverse problems: *(i)* strict propriety ensures that it is minimized only when the predictive distribution equals the true posterior, so it penalizes posterior collapse rather than rewarding it, *(ii)* its reduction to MAE for point-estimators places generative and regression models on a common footing, enabling fair comparison, and *(iii)* it is per-event and averaged, matching RMSE's granularity while measuring distributional rather than pointwise proximity to the truth.

4.2 Metric II: Spectrum Fidelity (Population-Level Accuracy)

CRPS does not measure the marginal, $p(z)$, across the dataset – the population-level quantity downstream analyses depend on. To evaluate this, we introduce a population-level diagnostic not targeted by existing SBI benchmarks (Lueckmann et al., 2021), based on the binned χ^2 statistic. Let $\{n_b^{\text{true}}\}$ and $\{n_b^{\text{pred}}\}$ denote the histogram bin counts of the true and predicted values of z over the test dataset. The spectrum fidelity is then defined as the Pearson χ^2 (Pearson, 1992):

$$\chi^2_{\text{spec}} = \sum_{b=1}^{B} \frac{\left(n_b^{\text{pred}} - n_b^{\text{true}}\right)^2}{n_b^{\text{true}}}, \tag{5}$$

where the sum runs over B bins, and assuming Poisson uncertainties with the variance in each bin approximated as $\sigma_b^2 \approx n_b^{\text{true}}$ which is valid when $n_b^{\text{true}} \gg 1$. For generative models, we construct the predicted histogram by using *one random posterior sample per event*. If the model is well calibrated, the resulting set of samples $\{\hat{z}_i \sim p(z \mid x_i)\}_{i=1}^{N}$ is marginally distributed as $p(z)$ by construction,[1] and, under the multinomial null, the expected χ^2_{spec} equals $B - 1$ (the number of degrees of freedom). For point-estimators, the single prediction per event is used directly.

[1]This follows from $p(z) = \int p(z \mid x)\, p(x)\, \mathrm{d}x$ where the integral is performed approximately by the sampling i.e. the test-set x_i are already draws from $p(x)$, which is never evaluated as a density.

We report the binned χ^2-metric as the primary spectrum fidelity diagnostic due to its interpretability and direct connection to goodness-of-fit tests standard in scientific analyses. The Earth Mover's distance (Rubner et al., 1998) is a natural unbinned alternative.

4.3 Metric III: Calibration (Uncertainty Trustworthiness)

CRPS assesses per-event distributional accuracy and spectrum fidelity assesses population-level marginal accuracy, but neither asks whether a model's stated uncertainty is reliable. Axis III therefore evaluates the empirical coverage, $C(\alpha)$ of predictive regions at nominal confidence level, α. A calibrated model produces a coverage curve tracking the diagonal $C(\alpha) \approx \alpha$; undercoverage ($C(\alpha) < \alpha$) indicates posteriors that are too narrow, overcoverage indicates posteriors that are too wide. The area between the curve and the diagonal, $\int_0^1 |C(\alpha) - \alpha| \, \mathrm{d}\alpha$, quantifies the overall deviation. Calibration should also be assessed *conditionally* on phase-space bins (e.g., event topology or signal-to-noise ratio), since good global coverage can hide systematic regional failures.

We use a conformal variant because it provides finite-sample guarantees and admits a cross-family comparison including models for which likelihood-based calibration diagnostics are unavailable (Araz & Spannowsky, 2025). Expected coverage curves are also standard in the SBI literature (Hermans et al., 2022; Talts et al., 2020; Lemos et al., 2023). Alternatives are discussed in appendix B. Following the forecasting principle of "maximize sharpness subject to calibration" (Gneiting et al., 2007), well-calibrated models should be ranked by sharpness (e.g., mean prediction-set width).

Point estimators without uncertainty cannot be assessed on this axis: conformalized residuals give constant-width intervals that cover trivially and carry no conditional information. Models should be ranked across families on Axes I and II, which admit every family, and read unavailability of Axis III as informative and a limitation worth reporting given the variance compression of section 3.2.

4.4 Non-Redundancy of the Axes

The three evaluation axes diagnose distinct failure modes, and none substitutes for another. In the idealized limit where a model attains the Bayes-optimal CRPS ($p_\theta(z \mid x) = p(z \mid x)$ everywhere), marginalization over x recovers the true spectrum $p(z)$ by construction. Away from this global optimum, however, empirical CRPS and spectrum fidelity χ^2_{spec} are non-redundant. Dataset-averaged CRPS is an unsigned, per-event distance statistic and does not retain information on where residual errors occur in x-space: errors with the same average CRPS can have very different dependence on x. In appendix C.6 we provide a controlled demonstration for two models with identical average CRPS but vastly different χ^2_{spec}. Conversely, population regularizers can force marginal agreement ($\chi^2_{\text{spec}} \approx 1$) through compensating per-event errors while scoring mediocre CRPS. Axis III (calibration) is orthogonal to both and tests whether the predictive uncertainty is calibrated: a model can achieve competitive CRPS and marginal shape while producing over- or under-confident posterior widths.

5 Benchmark I: A Synthetic Inverse Problem with a Known Multimodal Posterior

We first verify the mechanism of section 3.2 in a controlled setting where the true posterior is known analytically. Because the forward model is symmetric, conditional-mean collapse follows by construction.

Problem design We consider the forward model, $f : \mathbb{R} \to \mathbb{R}$:

$$x = f(z) = z^2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \tag{6}$$

with uniform prior $z \sim \pi(z) = \mathcal{U}(-a, a)$, where $a = 5$ and $\sigma_\varepsilon = 0.5$. The mapping $z \mapsto z^2$ is symmetric: for any observation $x > 0$, the values $\pm\sqrt{x}$ are equally likely *a priori*, producing an analytically tractable bimodal posterior (see appendix C). In this model, x controls a smooth transition from unimodal ($x \approx 0$) to strongly bimodal regimes, covering the generic many-to-one pathology. In this benchmark $p(z) = \pi(z)$ because z is drawn i.i.d. from the prior. In general, they differ and eq. (5) scores $p(z)$.

Table 2: Synthetic benchmark metrics (10,000 test events). Values are mean ± standard deviation across 5 random seeds. The regression wins on RMSE; the normalizing flow and MDN dominate all distributional metrics. Averaging the normalizing flow posterior recovers the regression's RMSE, CRPS, and spectral mismodeling. Calibration can only be assessed for uncertainty-aware architectures. χ^2 has 49 degrees of freedom; Best per column in **bold**.

Method	RMSE ↓	CRPS ↓	χ^2_{spec} (closest to 49)	Calibration deviance ↓
Regression MLP	$\mathbf{2.877}^*$	2.487	$(4.34 \pm 0.16) \times 10^5$	N/A
Heteroscedastic reg.	4.049 ± 0.037	1.509 ± 0.001	$(5.4 \pm 0.1) \times 10^3$	$\mathbf{0.0057 \pm 0.0001}$
MDN‡	4.062 ± 0.038	1.286 ± 0.001	$61.0 \pm 7.0^\dagger$	0.0109 ± 0.0001
Normalizing flow (NF)	4.091 ± 0.011	$\mathbf{1.285 \pm 0.002}$	$\mathbf{53.2 \pm 11.4}^\dagger$	0.0117 ± 0.0005
NF (posterior average)	2.879 ± 0.002	2.486 ± 0.002	$(1.82 \pm 0.02) \times 10^5$	N/A

* Lowest by construction (section 3.1); converged to constant prediction with negligible seed variance. † One random posterior sample for $\chi^2_{\text{spec}}/\text{ndf}$. ‡ Mixture density network needs manual initialization; see appendix C.3.

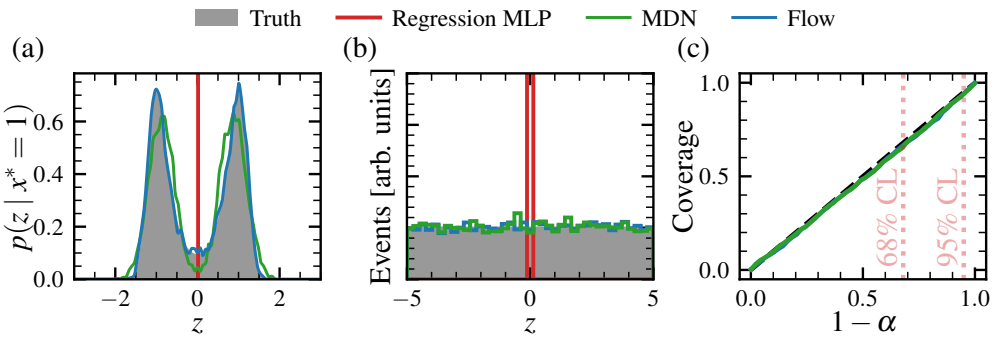


Figure 1: Synthetic benchmark results. **(a)** Per-event posteriors for an observation ($x^* = 1$) in bimodal regime. **(b)** Marginal distribution of reconstructed z over 10,000 test events. **(c)** Global conformal calibration coverage curves for the flow and MDN models.

Compared methods We compare four models on identical training data, differing only in their treatment of per-event uncertainty. The *regression MLP* is a 4-layer MLP trained with MSE; by posterior symmetry, its MSE-optimal prediction is $\hat{z}(x) = \mathbb{E}[z \mid x] = 0$ for all x, providing no per-event uncertainty and cannot be calibrated using conformal prediction. The *heteroscedastic regression* shares the backbone but outputs (μ, σ) trained by Gaussian NLL, modeling per-event uncertainty under a unimodal Gaussian assumption. The *mixture density network* (MDN) shares the regression backbone but outputs the parameters of a $K = 2$ Gaussian mixture, trained by NLL; this is our lightweight posterior-preserving baseline. The *conditional normalizing flow* is a 4-layer autoregressive rational-quadratic spline (ARQS) flow (Durkan et al., 2019) trained by NLL (Papamakarios et al., 2021), included as a more expressive reference. More details can be found in appendix C.

Evaluation setup Following section 4, we evaluate on 10,000 held-out events, drawing 500 posterior samples per event for the uncertainty-aware models (see appendix C.5), with $B = 50$ bins over $[-a, a]$ and a conformal split of 1,000 calibration and 9,000 evaluation events using the NLL as nonconformity score.

Results under pointwise metrics Table 2 (first column) confirms that the regression achieves the lowest RMSE, and averaging the flow's posterior samples per event recovers approximately the same value. The regression is simply the conditional mean of the flow's posterior.

Results under the proposed protocol The remaining columns of table 2 invert the ranking. The flow and MDN both dominate every distributional metric: CRPS scores are roughly half of the regression's, and $\chi^2_{\text{spec}}/\text{ndf} \approx 1$ indicates reconstructed marginal spectra statistically consistent with

Table 3: Physics benchmark for the $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ observable. Values are mean $\pm$ standard deviation across 4 random seeds, calibration deviance is single seed. Five methods evaluated under pointwise (RMSE) and distributional (CRPS, χ^2_{spec}) metrics. The transformer wins on RMSE; the discrete flow dominates distributional metrics. Best per column in **bold**.

Method	RMSE ↓	CRPS ↓	χ^2_{spec}/ndf (closest-to-1)	Calibration dev. ↓
Analytic	1.740	1.036	462 / 99	N/A
Transformer (MSE + MMD)	1.500 ± 0.011	0.936 ± 0.006	$(11{,}694 \pm 491) / 99$	N/A
Transformer (pure MSE)	$\mathbf{1.499 \pm 0.005}$	0.941 ± 0.003	$(14{,}085 \pm 760) / 99$	N/A
Discrete NF	1.766 ± 0.005	$\mathbf{0.544 \pm 0.001}$	$(266 \pm 44) / 99^{\dagger}$	**0.0238**
Continuous NF	1.780 ± 0.006	0.546 ± 0.001	$(\mathbf{252 \pm 29}) / \mathbf{99}^{\dagger}$	0.0886

† One random posterior sample for χ^2_{spec} / ndf. Extended table with more observables in appendix D.8.

the truth, compared to $\chi^2_{\text{spec}}/\text{ndf} \sim 10^4$ for the regression. Figure 1 makes the mechanism visible. At $x^* = 1$, the true posterior is bimodal at ± 1; the regression predicts $\hat{z} \approx 0$, a region of low posterior density, while the flow recovers both modes (panel a). Aggregated over the test set, the regression's marginal collapses to a spike at zero and the flow's matches the true uniform on $[-a, a]$ (panel b) – the extreme case of eq. (4). The heteroscedastic regression's χ^2_{spec} remains far above the MDN's despite per-event uncertainty – the unimodal Gaussian family cannot represent the true posterior.

All uncertainty-aware methods achieve nominal coverage (fig. 1(c), deviance < 0.02), but with very different sharpness: the heteroscedastic regression's 90% prediction sets span most of the prior, while the flow's and MDN's are disjoint, concentrated on the modes, and substantially narrower. Coverage alone is therefore insufficient – see table 2 for marginal deviances and appendix C.5 for prediction-set widths and conditional calibration.

Takeaways The same models on the same test data yield opposite rankings depending on the metric, but the deeper finding is that the relevant axis is posterior-family expressivity, not generative-vs-regression. A two-component mixture head on the regression backbone matches the flow, while a unimodal Gaussian head does not. This motivates including expressive flow-based models in Benchmark II where the posterior geometry is unknown a priori.

6 Benchmark II: A Real-World Multimodal Inverse Problem

We benchmark our approach on a particular particle physics inverse problem. Here, the scored quantity is a derived observable rather than the latent itself. The latent is $z \in \mathbb{R}^8$ (two four-momenta) and the observation x is a variable-length set of roughly nine tokens. The quantity scored is $\mathcal{O}(z) \in \mathbb{R}$ for $\mathcal{O} : \mathbb{R}^8 \to \mathbb{R}$, which is chosen because downstream inference depends on $\mathcal{O}(z)$ rather than on z. Concretely, the problem is reconstructing top-quark pairs in dileptonic decays. It exhibits all the pathologies targeted by the protocol: underdetermination (two neutrinos escape detection), combinatorial assignment, non-Gaussian posterior geometry, and measurement noise. Unlike Benchmark I, no analytic posterior is available and the evaluation instead uses the true latent z as a reference. Reconstructed spectra feed directly into top-mass measurements and new-particle searches (Favaro et al., 2025; CMS Collaboration, 2025b), with parameters of interest encoded in shapes and tails – exactly the regime section 3.2 warns about. Full physics background, observable definitions, and the forward model are given in appendix D.

Problem design We use a public Delphes (de Favereau et al., 2014) Monte Carlo dataset (Raine et al., 2023) (~950k simulated events with realistic detector effects) as our forward model. Inputs x are detector-level observables (jets, leptons, missing transverse energy); targets z are the top-quark four-momenta. Evaluation focuses on a multimodal derived observable $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$; its definition and additional observables are given in appendix D.3.

Compared methods We evaluate five methods under identical conditions. The *analytic solver* (Sonnenschein, 2005) algebraically imposes the kinematic mass constraints of the decay topology to return a single neutrino solution per event. The *transformer regression* (inspired by CMS Collab-

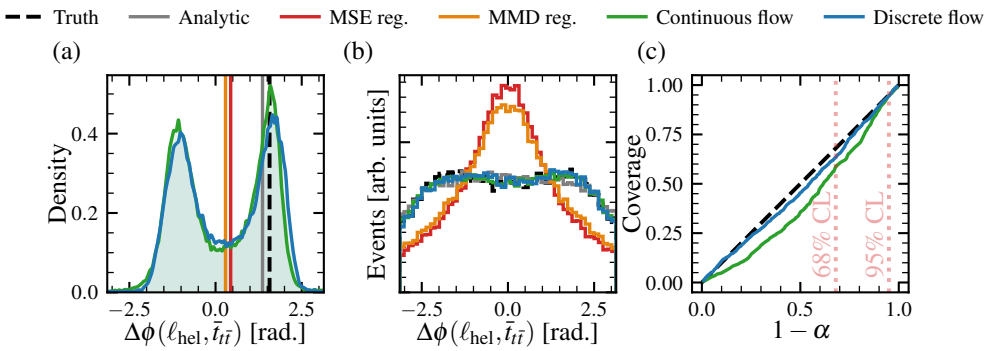


Figure 2: Top reconstruction benchmark results. **(a)** Reconstructed per-event posterior over $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$. The flow-based posteriors are nearly identical. **(b)** Marginal distribution of $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ over the test set. The flows sample a random point per event; the point estimators provide a point estimate without uncertainties. **(c)** Conformal coverage curve for the flows.

oration (2025a)) is a transformer encoder trained with a hybrid MSE + MMD loss; we also report a pure-MSE variant. The *discrete normalizing flow* (inspired by ν^2-flows (Raine et al., 2024)) is a conditional invertible network with exact likelihood. The *continuous normalizing flow* uses the conditional flow-matching objective (Chen et al., 2018; Lipman et al., 2023), with likelihood and sampling via ODE integration. Full details in appendix D.

Evaluation setup We follow the protocol of section 4, reporting CRPS, spectrum fidelity, and calibration coverage alongside RMSE for $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$. We draw 500 posterior samples per event from each flow; CRPS uses the full ensemble, χ^2_{spec} a single random draw per event. Calibration coverage is assessed for the flows via their log-likelihoods. Results are mean $\pm$ standard deviation across 4 random seeds per architecture; appendix D.7 shows that the rankings on χ^2_{spec} are unchanged under a sweep of the bin count and under the Earth Mover's distance.

Results under pointwise metrics Table 3 shows that the pure-MSE transformer achieves the best RMSE on $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$, with the MMD-regularized variant trailing slightly. The analytic solver's RMSE is comparable to the generative models rather than the transformers, because it commits to a single mode per event rather than computing a posterior mean: when the selected mode is correct, the error is small, when it is not, the error is of order the inter-mode separation. Both the solver and a single generative sample are "mode-picking" estimators; the transformers achieve lower RMSE precisely by averaging across modes, which is the source of the spectral distortion identified in section 3.2. The MMD transformer retains competitive RMSE and closes most of the marginal gap on unimodal observables (see appendix D), making it the natural choice under current evaluation practice (CMS Collaboration, 2025a). The generative models offer no obvious advantage on either axis.

Results under the proposed protocol The distributional metrics flip this conclusion. On $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$, the discrete flow achieves the best CRPS and lowest χ^2_{spec}, while the performance of both transformers remains orders of magnitude below that of the flow despite the MMD penalty. This pattern is exactly the prediction of section 3.2: the marginal penalty fixes what marginal information can fix – on unimodal observables, MMD reduces χ^2_{spec} by an order of magnitude and approaches that of the flows (appendix D) – but on the multimodal $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ it cannot correct the per-event conditional, and the spectral distortion persists (fig. 2(b)). Per-event posteriors (fig. 2(a)) make the mechanism visible: both flows capture the bimodal structure of $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ while both transformers collapse to the mean between modes, exactly like the synthetic pathology of fig. 1. The variance compression is visually clearest on unimodal observables (see appendix D.8).

The coverage curves (fig. 2(c)) expose a qualitative difference invisible to RMSE or CRPS: the discrete flow is well-calibrated with only slight undercoverage, while the continuous flow systematically undercovers despite comparable CRPS. The discrepancy possibly reflects both approximate ODE-based density evaluation and the mismatch between the flow-matching objective and the like-

lihood used for calibration – the discrete flow optimizes this likelihood directly. Two architectures indistinguishable on per-event accuracy thus produce qualitatively different uncertainty estimates, surfaced only by the full three-metric protocol.

Takeaways On a realistic problem, the protocol reverses the model selection that pointwise evaluation would make: the pure-MSE transformer wins on RMSE while the discrete flow dominates CRPS and spectrum fidelity, as predicted by section 3.2. Calibration further separates two flows indistinguishable under CRPS and χ^2_{spec}.

7 Discussion

Limitations Univariate CRPS loses joint posterior structure for high-dimensional z; multivariate proper scores (energy score (Gneiting et al., 2008); see appendix D.11) and joint coverage diagnostics (SBC (Talts et al., 2020), TARP (Lemos et al., 2023)) extend the protocol but the empirical scope of this work remains one-dimensional scored quantities. The multivariate form is preferable when the downstream measurement is a nonlinear function of several latent components, since matching each marginal separately does not constrain their joint dependence. The physics benchmark uses a single Monte Carlo generator with Delphes detector simulation, and we have not tested how the learned posteriors or their calibration behave under domain shift to full simulation or real collision data – validating this robustness is a prerequisite for experimental deployment. Finally, the Benchmark II models are trained on a modest dataset without exhaustive tuning, so the absolute values in table 3 should not be read as each architecture's best achievable performance. The ranking flip itself is robust: section 3.2 shows that no MMD strength can correct per-event multimodality, so while gaps may shrink with tuning, they cannot close.

Applicability beyond the studied benchmarks The protocol is domain-agnostic. Axes I and III apply to any many-to-one inverse problem with per-event uncertainty. Axis II requires the aggregated-latent setting of section 1, where the deliverable is a population marginal. For a single inversion of one observation, the protocol reduces to per-event accuracy and calibration. The protocol requires no domain-specific infrastructure: CRPS reduces to MAE for point-estimators, spectrum fidelity requires only a binned histogram, and calibration requires either a tractable likelihood or conformal nonconformity scores. The ranking flip is also not limited to bimodal posteriors: by eq. (3), any non-negligible within-posterior variance produces systematic tail compression that does not average away with more data, though the distortion is milder for heavy-tailed unimodal cases and negligible for sharply peaked ones. We recommend practitioners inspect posterior samples for a subset of events before defaulting to RMSE.

Open Problems Lightweight per-event uncertainty heads (e.g., MDNs) may correct spectral mismodeling more efficiently than population-level regularizers by directly modeling the within-posterior variance of eq. (3) (Cui et al., 2020). Whether they match flow-based fidelity on realistic multimodal problems is open. More broadly, distributional evaluation gains do not automatically translate into improved measurements, and how posterior samples should be carried through downstream inference remains largely unstudied.

8 Conclusion

Pointwise metrics are structurally misaligned with the scientific goals of supervised reconstruction: the law of total variance forces any point estimator converging to the conditional mean to produce a marginal spectrum narrower than the truth, a bias that does not diminish with more data. In the aggregated-latent setting this consequence is unmeasured, which is what our three-part protocol – CRPS, spectrum fidelity, and coverage-based calibration – addresses. A synthetic benchmark verifies the mechanism, and on a realistic particle-physics inverse problem model rankings reverse between pointwise and distributional evaluation while calibration separates architectures indistinguishable under CRPS. By rewarding posterior collapse, pointwise-only evaluation favors models whose spectra bias the measurements built on them. The protocol we propose makes this visible and gives practitioners a basis for comparison that scientific conclusions can rest on. Code and evaluation pipelines are linked in appendix E.

AI USE STATEMENT

In this work, we have used generative AI tools to draft parts of the paper, edit the text for readability, for polishing, and for reviewing parts of the manuscript for inconsistencies. The authors checked everything independently and made all final decisions. Additionally, we have used generative AI tools to aid documenting and formatting code. We have not used generative AI tools to generate or process data, to interpret results, or design the benchmarks. All AI-assisted text has been reviewed and edited by the authors, and all cited literature was checked against the original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only synthetic data and a publicly released simulated physics dataset (Raine et al., 2023). It involves no human subjects, personal data, or foreseeable dual-use risk. Its intended impact is on evaluation practice: pointwise-only evaluation can favor models whose reconstructed distributions bias downstream scientific conclusions.

REPRODUCIBILITY STATEMENT

The theoretical argument of section 3.2 follows from the law of total variance with assumptions stated inline. The synthetic forward model and analytic posterior of Benchmark I are given in eq. (6) and appendix C.1; data generation, architectures, training, and evaluation details are in appendices C.2 to C.4. Benchmark II uses a public dataset (Raine et al., 2023), with preprocessing, architectures, hyperparameters, and evaluation settings in appendices D.2 and D.5 to D.7 and robustness checks on binning and sampling conventions in appendix D.7. Anonymized code, configuration files, pinned dependencies, and notebooks reproducing every table and figure, together with pre-computed evaluation arrays, are linked in appendix E. Benchmark I results are averaged over 5 seeds and Benchmark II over 4 seeds; calibration deviance in Benchmark II is single-seed.

REFERENCES

Jack Y. Araz and Michael Spannowsky. Another fit bites the dust: Conformal prediction as a calibration standard for machine learning in high-energy physics, 2025. URL https://arxiv.org/abs/2512.17048.

Steve Baker and Robert D. Cousins. Clarification of the use of chi-square and likelihood functions in fits to histograms. *Nuclear Instruments and Methods in Physics Research*, 221(2):437–442, 1984. ISSN 0167-5087. doi: https://doi.org/10.1016/0167-5087(84)90016-4. URL https://www.sciencedirect.com/science/article/pii/0167508784900164.

Marco Bellagente, Anja Butter, Gregor Kasieczka, Tilman Plehn, Armand Rousselot, Ramon Winterhalder, Lynton Ardizzone, and Ullrich Köthe. Invertible networks or partons to detector and back again. *SciPost Phys.*, 9:074, 2020. doi: 10.21468/SciPostPhys.9.5.074. URL https://scipost.org/10.21468/SciPostPhys.9.5.074.

Christopher M. Bishop. Mixture density networks. WorkingPaper 4288, Aston University, 1994.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, January 2006. URL https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/.

Johann Brehmer, Pim de Haan, Sönke Behrends, and Taco Cohen. Geometric algebra transformer. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.18415.

Johann Brehmer, Víctor Bresó, Pim de Haan, Tilman Plehn, Huilin Qu, Jonas Spinner, and Jesse Thaler. A Lorentz-equivariant transformer for all of the LHC. *SciPost Phys.*, 19:108, 2025. doi: 10.21468/SciPostPhys.19.4.108. URL https://scipost.org/10.21468/SciPostPhys.19.4.108.

Ricky T. Q. Chen. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 2018.

The CMS Collaboration. Enhanced reconstruction of dileptonic top quark-antiquark events using supervised machine learning methods. Technical report, CERN, Geneva, 2025a. URL https://cds.cern.ch/record/2944724.

The CMS Collaboration. Observation of a pseudoscalar excess at the top quark pair production threshold. *Reports on Progress in Physics*, 88(8):087801, August 2025b. doi: 10.1088/1361-6633/adf7d3. URL https://doi.org/10.1088/1361-6633/adf7d3.

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. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1912789117.

Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated reliable regression using maximum mean discrepancy. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17164–17175. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c74c4bf0dad9cbae3d80faa054b7d8ca-Paper.pdf.

Maximilian Dax, Stephen R. Green, Jonathan Gair, Jakob H. Macke, Alessandra Buonanno, and Bernhard Schölkopf. Real-time gravitational wave science with neural posterior estimation. *Phys. Rev. Lett.*, 127:241103, December 2021. doi: 10.1103/PhysRevLett.127.241103. URL https://link.aps.org/doi/10.1103/PhysRevLett.127.241103.

Jerome de Favereau, Christophe Delaere, Pavel Demin, Andrea Giammanco, Vincent Lemaître, Alexandre Mertens, Michele Selvaggi, and The DELPHES 3 collaboration. Delphes 3: a modular framework for fast simulation of a generic collider experiment. *Journal of High Energy Physics*, 2014(2):57, 2014. doi: 10.1007/JHEP02(2014)057.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf.

Lukas Ehrke, John Andrew Raine, Knut Zoch, Manuel Guth, and Tobias Golling. Topological reconstruction of particle physics processes using graph neural networks. *Phys. Rev. D*, 107(11):116019, 2023. doi: 10.1103/PhysRevD.107.116019.

Luigi Favaro, Roman Kogler, Alexander Paasch, Sofia Palacios Schweitzer, Tilman Plehn, and Dennis Schwarz. How to unfold top decays. *SciPost Phys. Core*, 8:053, 2025. doi: 10.21468/SciPostPhysCore.8.3.053. URL https://scipost.org/10.21468/SciPostPhysCore.8.3.053.

Christopher A. T. Ferro, David S. Richardson, and Andreas P. Weigel. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15(1):19–24, 2008. doi: https://doi.org/10.1002/met.45. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.45.

Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B*, 69(2):243–268, 2007. URL https://EconPapers.repec.org/RePEc:bla:jorssb:v:69:y:2007:i:2:p:243-268.

Tilmann Gneiting, Larissa I. Stanberry, Eric P. Grimit, Leonhard Held, and Nicholas A. Johnson. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235, August 2008. ISSN 1863-8260. doi: 10.1007/s11749-008-0114-x. URL https://doi.org/10.1007/s11749-008-0114-x.

Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. OUP Oxford, 2020. ISBN 9780192586865.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017. URL http://arxiv.org/abs/1706.04599.

Joeri Hermans, Arnaud Delaunoy, François Rozet, Antoine Wehenkel, Volodimir Begy, and Gilles Louppe. A trust crisis in simulation-based inference? your posterior approximations can be unfaithful, 2022. URL https://arxiv.org/abs/2110.06581.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000. doi: 10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/wefo/15/5/1520-0434_2000_015_0559_dotcrp_2_0_co_2.xml.

Nathan Huetsch, Javier Mariño Villadamigo, Alexander Shmakov, Sascha Diefenbacher, Vinicius Mikuni, Theo Heimel, Michael Fenton, Kevin Greif, Benjamin Nachman, Daniel Whiteson, Anja Butter, and Tilman Plehn. The landscape of unfolding with machine learning. *SciPost Phys.*, 18:070, 2025. doi: 10.21468/SciPostPhys.18.2.070. URL https://scipost.org/10.21468/SciPostPhys.18.2.070.

Alexander Jordan, Fabian Krüger, and Sebastian Lerch. Evaluating probabilistic forecasts with scoringrules. *Journal of Statistical Software*, 90(12):1–37, 2019. doi: 10.18637/jss.v090.i12. URL https://www.jstatsoft.org/index.php/jss/article/view/v090i12.

Yuji Kim and Nori Nakata. Geophysical inversion versus machine learning in inverse problems. *Leading Edge*, 37(12):894–901, December 2018. doi: 10.1190/tle37120894.1.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *CoRR*, abs/1807.00263, 2018. URL http://arxiv.org/abs/1807.00263.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19256–19273. PMLR, July 2023. URL https://proceedings.mlr.press/v202/lemos23a.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. URL http://arxiv.org/abs/1711.05101.

Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 343–351. PMLR, May 2021. URL https://proceedings.mlr.press/v130/lueckmann21a.html.

George Papamakarios and Iain Murray. Fast ϵ -free inference of simulation models with bayesian conditional density estimation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL http://jmlr.org/papers/v22/19-1028.html.

Karl Pearson. *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling*, pp. 11–28. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_2. URL https://doi.org/10.1007/978-1-4612-4380-9_2.

Antoine Petitjean, Anja Butter, Kevin Greif, Sofia Palacios Schweitzer, Tilman Plehn, Jonas Spinner, and Daniel Whiteson. Generative unfolding of jets and their substructure, 2025. URL https://arxiv.org/abs/2510.19906.

John Andrew Raine, Matthew Leigh, Knut Zoch, Lukas Ehrke, Debajyoti Sengupta, and Tobias Golling. Dileptonic ttbar neutrino regression dataset, July 2023. URL https://doi.org/10.5281/zenodo.8113516.

John Andrew Raine, Matthew Leigh, Knut Zoch, and Tobias Golling. Fast and improved neutrino reconstruction in multineutrino final states with conditional normalizing flows. *Phys. Rev. D*, 109:012005, January 2024. doi: 10.1103/PhysRevD.109.012005. URL https://link.aps.org/doi/10.1103/PhysRevD.109.012005.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 59–66, January 1998. doi: 10.1109/ICCV.1998.710701.

Hongyu Shen, E A Huerta, Eamonn O'Shea, Prayush Kumar, and Zhizhen Zhao. Statistically-informed deep learning for gravitational wave parameter estimation. *Machine Learning: Science and Technology*, 3(1):015007, November 2021. doi: 10.1088/2632-2153/ac3843. URL https://doi.org/10.1088/2632-2153/ac3843.

Alexander Shmakov, Michael James Fenton, Ta-Wei Ho, Shih-Chieh Hsu, Daniel Whiteson, and Pierre Baldi. SPANet: Generalized permutationless set assignment for particle physics using symmetry preserving attention. *SciPost Phys.*, 12(5):178, 2022. doi: 10.21468/SciPostPhys.12.5.178.

Emil Y. Sidky and Xiaochuan Pan. Report on the aapm deep-learning sparse-view ct grand challenge. *Medical Physics*, 49(8):4935–4943, 2022. doi: https://doi.org/10.1002/mp.15489. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.15489.

Lars Sonnenschein. Algebraic approach to solve $t\bar{t}$ dilepton equations. *Phys. Rev. D*, 72:095020, November 2005. doi: 10.1103/PhysRevD.72.095020. URL https://link.aps.org/doi/10.1103/PhysRevD.72.095020.

Jonas Spinner, Victor Bresó, Pim De Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer. Lorentz-equivariant geometric algebra transformers for high-energy physics. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2405.14806.

Vincent Stimper, David Liu, Andrew Campbell, Vincent Berenz, Lukas Ryll, Bernhard Schölkopf, and José Miguel Hernández-Lobato. normflows: A pytorch package for normalizing flows. *Journal of Open Source Software*, 8(86):5361, 2023. doi: 10.21105/joss.05361. URL https://doi.org/10.21105/joss.05361.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration, 2020. URL https://arxiv.org/abs/1804.06788.

Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, and Auroop R. Ganguly. Generating high resolution climate change projections through single image super-resolution: an abridged version. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 5389–5393. AAAI Press, 2018. ISBN 9780999241127.

Zhendong Wang, Ruijiang Gao, Mingzhang Yin, Mingyuan Zhou, and David Blei. Probabilistic conformal prediction using conditional random samples. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8814–8836. PMLR, April 2023. URL https://proceedings.mlr.press/v206/wang23n.html.

Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2021. doi: 10.1109/TPAMI.2020.2982166.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Borislavov Kovachki, Zachary E Ross, Katherine Bouman, and Yisong Yue. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=U3PBITXNG6.

A Evaluation Metrics

A.1 Empirical CRPS estimator

For a predictive distribution represented by N posterior samples $\{\hat{z}^{(k)}\}_{k=1}^{N}$, the CRPS of eq. (1) is estimated as

$$\widehat{\mathrm{CRPS}} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{z}^{(k)} - z\right| - \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{j=1}^{N}\left|\hat{z}^{(k)} - \hat{z}^{(j)}\right|, \tag{7}$$

computable in $\mathcal{O}(N \log N)$ via sorting (Jordan et al., 2019). The estimator has finite-sample bias of $\mathcal{O}(1/N)$ from the second term (Ferro et al., 2008). For the sample sizes used in this work ($N = 500$ for both benchmarks) the bias $\approx 0.05\%$ of the reported CRPS values is well below the gaps between methods. Reducing F to a Dirac distribution at a point prediction $\hat{z}$ recovers $\widehat{\mathrm{CRPS}} = |\hat{z} - z|$, the MAE.

A.2 Spectrum fidelity

The binned χ^2 statistic of eq. (5) requires only that bin counts are well-approximated as Poisson, i.e. $n_b^{\text{true}} \gtrsim 10$. Bin counts and per-benchmark binning choices are stated in the respective evaluation subsections (appendices C and D). For tail-heavy spectra where some bins violate the Poisson approximation, a likelihood-ratio statistic such as Baker-Cousins (Baker & Cousins, 1984) is a more accurate goodness-of-fit measure. We discuss this where relevant in appendix D.8.

A.3 Calibration

The conformal procedure used throughout this work, including the choice of nonconformity score per model and the finite-sample marginal coverage guarantee, is described next.

B Alternative Calibration Procedures

The protocol of section 4.3 uses conformal calibration as a concrete instantiation of the calibration axis, but the protocol's requirement is that calibration is assessed – not which procedure is used. We briefly review three methods and indicate when each is appropriate.

Conformal calibration Given a nonconformity score $s(z, x)$ (e.g. the NLL), the threshold $\hat{q}_{1-\alpha}$ is the $\lceil (1-\alpha)(n_{\text{cal}}+1) \rceil$-th smallest score on a held-out calibration set (Araz & Spannowsky, 2025). The resulting prediction set satisfies $\mathbb{P}(z \in \mathcal{C}_{1-\alpha}(x)) \geq 1-\alpha$ under exchangeability, giving a finite-sample marginal coverage guarantee that holds across model families. We use it throughout this work for cross-family comparability.

HPD credible intervals For models with tractable densities, the highest posterior density region $\mathcal{R}_\alpha = \{z : p_\theta(z \mid x) \geq \tau_\alpha\}$ tests the learned density directly without requiring a calibration split. It offers no finite-sample guarantee and is most appropriate when the density is the object of interest itself, e.g. for likelihood-trained models being assessed on their own training objective.

PIT / CDF-based diagnostics The probability integral transform of the true value under the predicted CDF should be uniform under correct calibration (Gneiting et al., 2007) and deviations reveal systematic biases. PIT is most informative for univariate predictions and provides a richer diagnostic than scalar coverage, but no formal guarantee.

SBC and TARP For joint posteriors over high-dimensional latents, simulation-based calibration (Talts et al., 2020) and TARP (Lemos et al., 2023) extend coverage diagnostics to the joint setting. They are the natural choice when the protocol's univariate CRPS is replaced by the energy score (section 4.1) for joint-posterior applications.

Choosing among them We recommend to use conformal prediction when comparing across model families (its finite-sample guarantee is family-agnostic). Use HPD or PIT when assessing a single model's density quality on its own terms. Use SBC or TARP when the latent of interest is high-dimensional and the joint posterior structure matters.

C Benchmark I: Synthetic Inverse Problem

This appendix provides architecture, training, and evaluation details for the synthetic benchmark of section 5, along with extended results supporting the main-text claims.

C.1 Forward Model

We want $p(z \mid x)$ for the model $x = z^2 + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $z \sim \mathcal{U}(-a, a)$. Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the standard normal distribution with mean μ and variance σ^2, and $\mathcal{U}(-a, a)$ denotes the uniform distribution on the range $[-a, a]$. From Bayes' theorem:

$$p(z \mid x) = \frac{p(x \mid z)\, p(z)}{p(x)}. \tag{8}$$

For a given z we have $x \mid z \sim \mathcal{N}(z^2, \sigma_\varepsilon^2)$ which is expanded to

$$p(x \mid z) = \frac{1}{\sqrt{2\pi}\,\sigma_\varepsilon} \exp\left(-\frac{(x-z^2)^2}{2\sigma_\varepsilon^2}\right).$$

The prior $\pi(z)$ is uniform, so it contributes no z dependence on the likelihood. The unnormalized posterior is:

$$p(z \mid x) \propto p(x \mid z)\, p(z) \propto \exp\left(-\frac{(x-z^2)^2}{2\sigma_\varepsilon^2}\right) \quad \text{for } z \in [-a, a] \tag{9}$$

The factors $\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}$ and $\frac{1}{2a}$ are constants w.r.t. z and are absorbed into the normalization. This leads to the unnormalized log-posterior:

$$\log \tilde{p}(z \mid x) = -\frac{1}{2}\left(\frac{x-z^2}{\sigma_\varepsilon}\right)^2, \quad \text{for } z \in [-a, a] \tag{10}$$

$p(x)$ is intractable analytically:

$$p(x) = \int_{-a}^{a} p(x \mid z)\, p(z)\, \mathrm{d}z = \frac{1}{2a} \int_{-a}^{a} \frac{1}{\sqrt{2\pi}\,\sigma_\varepsilon} \exp\left(-\frac{(x-z^2)^2}{2\sigma_\varepsilon^2}\right) \mathrm{d}z, \tag{11}$$

Table 4: Benchmark I model comparison. All non-flow models share a 4-layer MLP backbone and differ only in output head and training objective.

Model	Backbone	Parameters	Output	Loss/Objective	Posterior
Regression	4-layer MLP	49,921	$\hat{z} \approx \mathbb{E}[z \mid x]$	MSE	$\delta(\hat{z})$
Heteroscedastic	4-layer MLP	50,050	$(\mu, \log \sigma)$	Gaussian NLL	$\mathcal{N}(\mu, \sigma^2)$
MDN	4-layer MLP	50,566	$\{\pi_k, \mu_k, \log \sigma_k\}_{k=1}^{2}$	Mixture NLL	$\sum_k \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$
Normalizing flow	4-layer ARQS	74,596	flow parameters	NLL	$p_\theta(z \mid x)$

but constant with respect to z. We can evaluate the unnormalized posterior on a grid and normalize it numerically:

$$p(z \mid x) = \frac{\exp(\log p)}{\int_{-a}^{a} \exp(\log p)\, \mathrm{d}z}. \tag{12}$$

We use eq. (12) as the truth posterior to compare against in Benchmark I, providing an exact ground-truth posterior for every observation, x.

C.2 Data generation and splits

We sample $z \sim \mathcal{U}(-a, a)$ with $a = 5$, compute $y = z^2 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $\sigma_\varepsilon = 0.5$, and split independently-drawn samples into training (50,000), validation (10,000), and test (10,000) sets. Splits are generated with fixed random seeds; because the prior and noise are sampled fresh for each set, train-test contamination is impossible by construction.

C.3 Model architectures and training details

The model architectures used in Benchmark I are shown in table 4. Reference implementations of all four models, training scripts, and the exact configuration files used to produce table 2 and fig. 1 can be found in the linked repository (appendix E). All models are trained with AdamW (Loshchilov & Hutter, 2017)(learning rate 10^{-3}, batch size 512) on 50,000 samples for 80 epochs to ensure all models are fully converged. We do not employ early stopping, and we use cosine annealing learning rate scheduling. We have trained on a `NVIDIA A100 80GB PCIe` GPU. Training the MLP-backbone models takes approximately one minute, while training the normalizing flow takes approximately 7 minutes.

Shared backbone The regression, heteroscedastic regression, and MDN share a 4-layer fully-connected backbone with 128 hidden units per layer, ReLU activations, and a single scalar input x. Models differ only in the output head and training objective; the normalizing flow uses a different architecture entirely (described below).

MDN initialization On the symmetric posterior of $x = z^2 + \varepsilon$, MDNs with default random initialization frequently collapse to an effectively unimodal posterior – a known property of MDN training on symmetric multimodal targets (Bishop, 1994). We break the symmetry by initializing the two component means at $\mu_1 = +1$ and $\mu_2 = -1$, with no other changes to architecture, optimizer, or training schedule. This stabilizes training across all 5 seeds reported in table 2. This instability further motivates using more expressive and stabler flows in Benchmark II.

Conditional normalizing flow We use a conditional normalizing flow built with the `normflows` library (Stimper et al., 2023). The base distribution is a standard 1D Gaussian (fixed mean and unit variance). The flow stacks 4 transform blocks, each consisting of an autoregressive rational-quadratic spline (ARQS) coupling (Durkan et al., 2019) followed by an LU-decomposed linear permutation. Each spline transform has 2 hidden blocks with 64 hidden units conditioning on the scalar observation x (1 context channel). Default `normflows` hyperparameters are used for spline tail bounds and number of knots.

Table 5: Expanded version of table 2. Synthetic benchmark metrics (10,000 test events). Values are mean ± standard deviation across 5 random seeds. The regression wins on RMSE; the normalizing flow and MDN dominate all distributional metrics. Averaging the normalizing flow posterior recovers the regression's RMSE, CRPS, and spectral mismodeling. Calibration only possible for uncertainty-aware architectures. χ^2 has 49 degrees of freedom; Best (mean) per column in **bold**.

Method	RMSE ↓	CRPS ↓	χ^2_{spec} (closest to 49)	Cal. dev.($\times 10^{-2}$) ↓
Regression MLP	$\mathbf{2.877}^*$	2.487	$(4.34 \pm 0.16) \times 10^5$	—
Regression + MMD	2.958 ± 0.007	2.587 ± 0.002	$(2.830 \pm 0.004) \times 10^4$	—
Heteroscedastic reg.	4.049 ± 0.037	1.509 ± 0.001	$(5.4 \pm 0.1) \times 10^3$	$\mathbf{0.0057 \pm 0.0001}$
MDN‡	4.062 ± 0.038	1.286 ± 0.001	$61.0 \pm 7.0^\dagger$	0.0109 ± 0.0001
MDN (posterior mean)	2.879 ± 0.014	2.486 ± 0.001	$(1.88 \pm 0.01) \times 10^5$	—
Normalizing flow (NF)	4.091 ± 0.011	$\mathbf{1.285 \pm 0.002}$	$\mathbf{53.2 \pm 11.4}^\dagger$	0.0117 ± 0.0005
NF (posterior average)	2.879 ± 0.002	2.486 ± 0.002	$(1.82 \pm 0.02) \times 10^5$	—

* Lowest by construction (section 3.1); converged to constant prediction with negligible seed variance. † One random posterior sample for χ^2_{spec}/ndf. ‡ Mixture density network needs manual initialization.

C.4 Evaluation details

The regression MLP model reports a point estimate per event, and CRPS reduces to MAE for this model. For the remaining models, CRPS is calculated from $N = 500$ posterior samples per event using the sorting-based estimator of eq. (7). The estimator bias from finite N is $\mathcal{O}(1/N) \approx 0.05\%$ (Ferro et al., 2008), well below the gaps reported in table 2.

With $B = 50$ uniform bins over $[-5, 5]$, expected counts under the true uniform prior $\pi(z)$ are ~ 200 per bin, validating the Gaussian approximation $\sigma_b^2 \approx n_b^{\text{true}}$. The regression's collapse to $z \approx 0$ concentrates predictions in a single bin, producing the dominant contribution to χ^2_{spec}.

We apply conformal prediction as a model-agnostic calibration diagnostic. For the uncertainty-aware models, the nonconformity score is the NLL $s(z, x) = -\log p_\theta(z \mid x)$. The regression has no notion of nonconformity score that scales with input – a uniform residual across x produces calibrated coverage trivially via fixed-width intervals.

Given a calibration set of size n_{cal}, the conformal threshold is then given as the empirical threshold on the calibration set corresponding to some mis-coverage level, α (e.g., 0.05 for 95% confidence) (Araz & Spannowsky, 2025):

$$\hat{q}_{1-\alpha} = \text{the } \lceil (1-\alpha)(n_{\text{cal}} + 1) \rceil\text{-th smallest value of } \{s_i\}. \tag{13}$$

A well-calibrated model should produce a coverage curve that tracks the diagonal.

Calibration alone is not enough – a model can be perfectly calibrated yet produce sets so wide that they are practically useless. We therefore also report the average Lebesgue size of the conformal prediction set:

$$\hat{C}_{1-\alpha}(y) \;=\; \{\, z : s(z, y) \le \hat{q}_{1-\alpha} \,\}$$

on a uniform grid of 1,000 points over $[-a, a]$. The reported width is the total Lebesgue measure (sum of interval lengths for disjoint regions).

C.5 Additional Results

This subsection presents extended visualizations and a sensitivity analysis supporting the main-text claims of section 5.

Per-event posteriors across the unimodal-to-bimodal transition Figure 3 extends fig. 1(a) by showing the recovered per-event posteriors at six different observations x^* spanning the unimodal regime ($x^* \approx 0$, where the two modes merge) through the strongly bimodal regime ($x^* \gg 1$ where modes are well separated at $\pm\sqrt{x^*}$). The flow and MDN track the analytic posterior across all regimes, including the smooth transition between regimes. The heteroscedastic regression – shown here but not in the main text – collapses to a single Gaussian centered near zero with inflated variance whenever

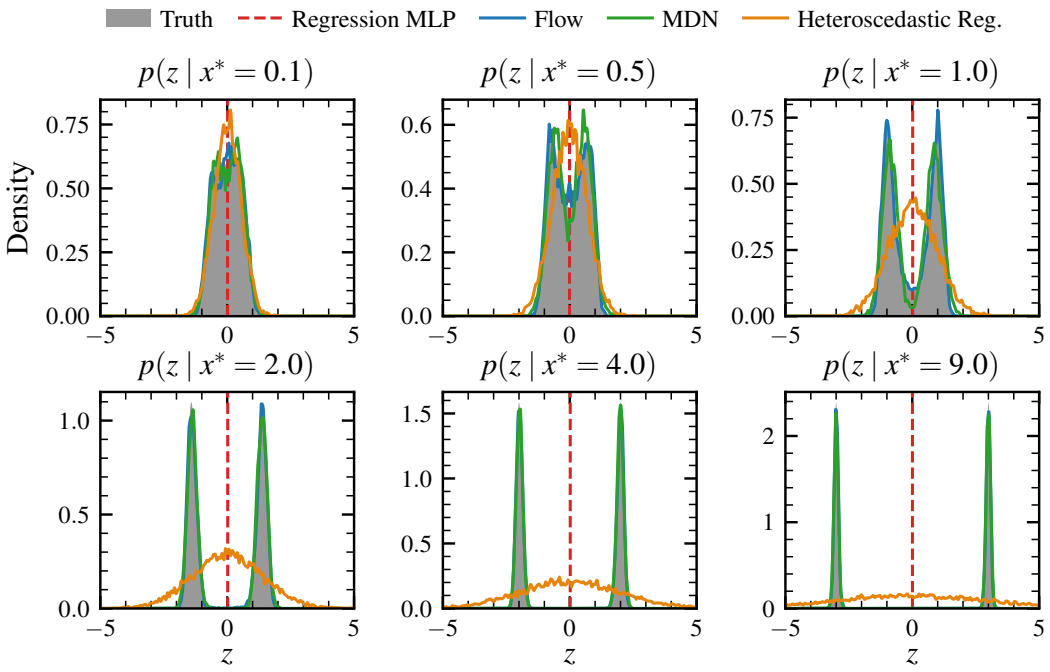


Figure 3: Additional conditional posteriors for the toy model presented in section 5. We present the reconstructed posteriors for six different x^* values similar to fig. 1(a). Here, we also include the heteroscedastic regression model.

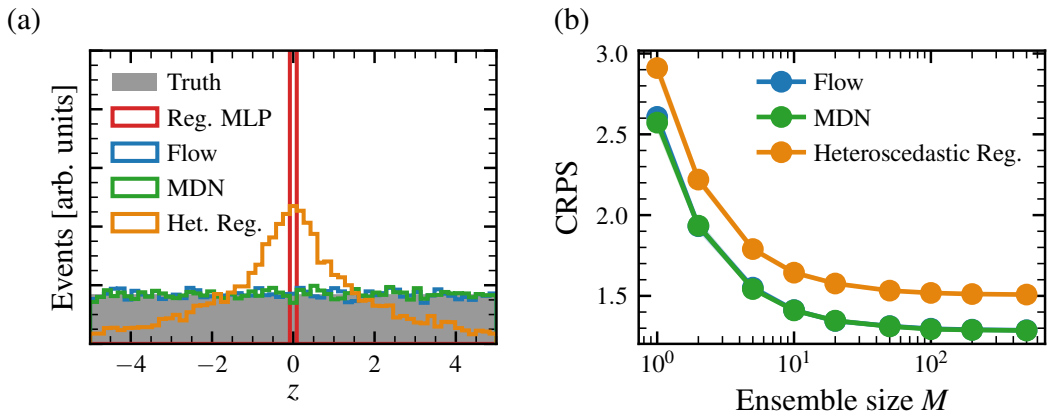


Figure 4: **(a)** Marginal z over 10,000 test events. Similar to fig. 1(b), but we have included the marginal recovered by the heteroscedastic regression. **(b)** The sensitivity of the CRPS score is measured by calculating it as a function of ensemble size, M, for the distributional models. The flow's curve coincides with MDN's.

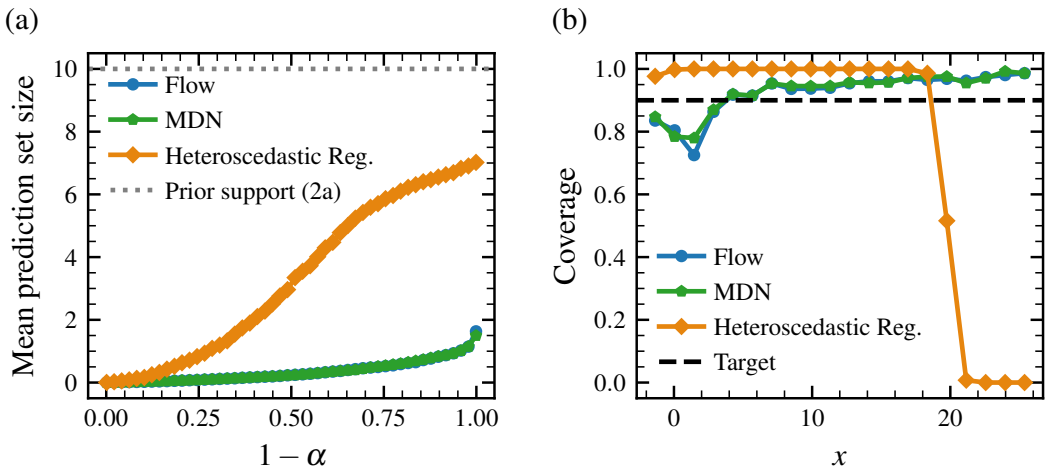


Figure 5: Calibration diagnostics beyond marginal coverage. **(a)** Mean conformal prediction-set size at each nominal coverage level: all methods are calibrated (fig. 1(c)) but produce sets of very different widths, demonstrating that coverage and sharpness are independent diagnostics. **(b)** Empirical coverage at the 90% nominal level as a function of the observation x, showing whether conditional calibration is uniform across phase space or hides regional deviations.

the true posterior is bimodal, illustrating the failure mode quantified by its χ^2_{spec} in table 2. Per-event uncertainty is insufficient if the predicted family cannot represent the true posterior geometry.

Marginal recovery including the heteroscedastic regression Figure 4(a) extends fig. 1(b) with the heteroscedastic regression's marginal. The flow and MDN marginals match the true uniform on $[-a, a]$ and the regression collapses to a spike at zero. The heteroscedastic regression fall somewhere in between, and it overestimates the center and underestimates the tails. This visually confirms the variance compression argument (section 3.2) that adding per-event uncertainty improves the marginal over a pure point estimator but cannot recover bimodal structure under a unimodal posterior assumption.

CRPS sensitivity to posterior ensemble size Figure 4(b) shows dataset-averaged CRPS as a function of the number of posterior samples M used for its estimation, for each distributional model. The estimator stabilizes well below the per-method gaps reported in table 2: by $M \approx 100$ all curves are within Monte Carlo noise of their asymptotic values, justifying the choice of $M = 500$ in the main text and confirming that method rankings are robust to ensemble size. The estimator bias from finite N is $\mathcal{O}(1/N) \approx 0.05\%$ (Ferro et al., 2008) when $M = 500$.

Prediction-set efficiency Figure 5(a) shows the mean prediction-set size as a function of nominal coverage $1 - \alpha$ for each model calibrated with conformal prediction. All three uncertainty-aware methods achieve nominal coverage (see fig. 1(c)), but their prediction sets differ substantially in size. At 90% nominal coverage, the heteroscedastic regression's sets span more than half of the full prior support $[-a, a]$ and convey almost no information beyond the prior. In contrast, the MDN and flow produce sets a few times narrower, concentrated as two disjoint intervals around the true modes $\pm\sqrt{x}$ (fig. 3). This quantifies the claim of section 4 that coverage alone is an insufficient calibration diagnostic: a model can be perfectly calibrated and yet uninformative, and sharpness must be reported alongside coverage to distinguish and rank the two.

Conditional calibration across the posterior-geometry transition Figure 5(b) shows empirical coverage at the 90% nominal level as a function of the observation x, probing whether marginal calibration hides regions of systematic over- or undercoverage. This addresses the recommendation of section 4.3 that calibration should be assessed conditionally in bins of phase space rather than only marginally. In addition to marginal calibration, we show the conditional calibration in fig. 5(b) at 90%. The heteroscedastic model achieves marginal calibration by overcovering for small x and

undercovering for large x. Only conditional calibration makes this visible. The MDN and flow slightly undercover in the unimodal regime and slightly overcover in the strongly bimodal regime.

C.6 Non-Redundancy via Matched-CRPS Predictors

To demonstrate that dataset-averaged CRPS and spectrum fidelity χ^2_{spec} diagnose distinct failure modes, we construct two synthetic predictors in the Benchmark I setting, both starting from the exact analytic posterior $p(z \mid x)$:

- Model A (x-independent error): Inflates posterior draws with additive Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma_A^2)$ independently of x.
- Model B (x-correlated error): Introduces a systematic shift proportional to the observation, $\hat{z} = z + c \cdot x$.

The parameter c in Model B is tuned so that its dataset-averaged CRPS matches that of Model A to within sample noise. Despite matched per-event distributional scores, their marginal fidelity differ substantially as shown in table 6.

Table 6: Demonstration of non-redundancy between per-event distributional accuracy (CRPS) and spectrum fidelity (χ^2_{spec}) in the Benchmark I setting. Both models start from the exact posterior and are tuned to match average CRPS within sample noise; the structure of the residual error leads to starkly different marginal fidelity.

Predictive Family	**Avg. CRPS** $\downarrow$	$\boldsymbol{\chi^2_{\text{spec}}}$
Model A (x-independent error)	0.74785	77 543.70
Model B (x-correlated error)	0.74785	30 739.93

Thus, the scalar average over CRPS (Axis I) does not reveal whether residual errors are structured with respect to x, while spectrum fidelity (Axis II) measures the consequences of those errors after aggregation. Reporting both metrics therefore provides information that neither metric can recover alone.

D Benchmark II: Dileptonic Top-Quark Reconstruction

This appendix provides architecture, training, and evaluation details for the physics benchmark of section 6, along with extended results supporting the main-text claims.

D.1 Physics background

In this work, we evaluate our method on the reconstruction of top quark pairs $(t\bar{t})$ produced in proton-proton collisions. The top quark is the heaviest known elementary particle, and its precise reconstruction is a cornerstone of the Large Hadron Collider (LHC) physics program.

D.1.1 Dileptonic Decay Topology

Top quarks decay almost exclusively into a W boson and a bottom (b) quark. When both W bosons from a $t\bar{t}$ pair decay into a charged lepton $(\ell = e, \mu)$ and a neutrino (ν), the process is called the *dileptonic decay channel*. The full decay chain is represented as:

$$pp \to t\bar{t} \to \left(W^+ b\right)\left(W^- \bar{b}\right) \to \left(\ell^+ \nu_\ell b\right)\left(\ell^- \bar{\nu}_\ell \bar{b}\right). \tag{14}$$

The experimental signature in the detector consists of two high-momentum leptons, two jets originating from the b-quarks (b-jets), and significant Missing Transverse Energy $\left(\vec{E}_T^{miss} \text{ or MET}\right)$ due to the undetected neutrinos.

From an algorithmic perspective, the reconstruction of the parent top quark four-momenta presents three primary challenges:

1. **Underdetermined Kinematics:** While the detector measures the transverse components of the sum of the neutrino momenta $\left(\vec{E}_T^{miss}\right)$, the individual longitudinal momenta (p_z) and the specific distribution of transverse momentum between the two neutrinos are unknown. This results in a system with six unknown degrees of freedom (the three-momentum components for each neutrino) but only two direct experimental constraints. Even when assuming the known masses of the W boson and the top quark as constraints, the system often remains underdetermined or yields multiple algebraic solutions.
2. **Combinatorial Ambiguity:** In a standard event, the detector identifies two b-jets, but it is not a priori known which jet originated from the top quark and which from the anti-top quark. For n additional light-flavor jets in the event, the number of possible permutations for the final-state assignment grows factorially, creating a complex assignment problem.
3. **Detector Resolution and Noise:** The measured momenta of jets and the $\vec{E}_T^{miss}$ are subject to experimental uncertainties and resolution effects. Traditional analytical "kinematic fitting" methods often fail when the measured values fluctuate such that no physical solution exists for the mass constraints.

D.2 Dataset details

We use the public Delphes (de Favereau et al., 2014) Monte Carlo dataset of Raine et al. (2023) (Zenodo DOI 10.5281/zenodo.8113516), released under CC-BY 4.0. The dataset comprises ∼950k simulated dileptonic $t\bar{t}$ events with realistic detector effects via Delphes. Event generation, selection, and detector simulation details are in the dataset release and Raine et al. (2024). We use the dataset's provided train/test splits without modification and we only consider the events generated with MadGraph.

The dataset was originally constructed for neutrino reconstruction, with bundled targets (called `neutrinos`) corresponding to the two neutrino four-momenta. For top-quark reconstruction we instead construct the target vectors ourselves from the dataset key `truth_particles`. The latent target is two four-vectors per event. The observables per event comprise up to six four-vector jets, two lepton four-vectors, a missing-transverse-energy vector (p_T and ϕ), and a boolean array indicating whether jets are b-tagged (whether a reconstructed jet is *likely* originating from a bottom quark).

D.3 Observable definitions

Once the four-momenta of the top quark (p_t) and the anti-top quark $(p_{\bar{t}})$ are reconstructed and the inverse problem is solved, several high-level observables can be computed for the actual physical study of the process. These variables are critical for testing Standard Model predictions and searching for new physics. The distributions over the observables we use here, vary from low-variance and unimodal $\left(p_{\mathrm{T}}^{t\bar{t}}\right)$ to higher-variance and multimodal $(\Delta\phi(\ell_{\mathrm{hel}},\, \bar{t}_{t\bar{t}}))$.

D.3.1 Global Kinematic Variables

These observables describe the overall scale and motion of the $t\bar{t}$ system.

- **Invariant Mass $(m_{t\bar{t}})$:** Defined as the Lorentzian norm of the combined four-momentum, $m_{t\bar{t}} = \sqrt{(p_t + p_{\bar{t}})^2}$. It represents the total energy scale of the $t\bar{t}$ production. In the context of machine learning, this is often the most important target for reconstruction, as local deviations in the $m_{t\bar{t}}$ spectrum could indicate the existence of heavy resonant particles (e.g., a heavy Higgs boson) decaying into top pairs.
- **Transverse Momentum $\left(p_{\mathrm{T}}^{t\bar{t}}\right)$:** The magnitude of the vector sum of the transverse momenta, $p_{\mathrm{T}}^{t\bar{t}} = |\vec{p}_{T,t} + \vec{p}_{T,\bar{t}}|$. In a simple leading-order collision, the top quarks are produced back-to-back in the transverse plane, leading to $p_{\mathrm{T}}^{t\bar{t}} \approx 0$. A non-zero value typically indicates the presence of "Initial State Radiation" (ISR), where the incoming partons radiate gluons before colliding.

D.3.2 ANGULAR AND SPIN-SENSITIVE OBSERVABLES

These variables describe the relative geometry of the decay products and are sensitive to the quantum state (spin) of the quarks.

- **Azimuthal Angle Difference** ($\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$)**:** The angle between the lepton and the anti-top quark in the plane transverse to the beam pipe in the helicity frame of the $t\bar{t}$ system, $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}}) = |\phi_{\ell_{\text{hel}}} - \phi_{\bar{t}_{t\bar{t}}}|$. This metric is a close proxy for the spin correlation between the top and the anti-top quarks in a pair system, see the next paragraph for c_{hel}.
- **Spin Correlation Variable** (c_{hel}): The spin correlation observable c_{hel} is computed from the scalar product of the lepton momentum vectors in the rest frames of their respective parent top quarks. It is closely related to $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ and the equivalent angle between the lepton and its parent top quark. It is a suitable variable to distinguish the spin state of the top quark pair and to search for top quark pairs which were produced by heavier (pseudo-)scalar particles.
  To compute c_{hel}, one has to boost the lepton momenta into the $t\bar{t}$ rest frame and then into the parent top quark frames. In this frame, c_{hel} is

$$c_{\text{hel}} = \hat{\ell}^1 \cdot \hat{\ell}^2 \tag{15}$$

D.4 FORWARD MODEL

The forward model maps generator-level top-quark four-momenta $\boldsymbol{z}$ to detector-level observations $\boldsymbol{x}$ through three stochastic stages: *(i)* hard-scattering matrix-element generation, *(ii)* parton showering and hadronization, and *(iii)* detector response. The Raine et al. (2023) dataset uses MadGraph (v3.10) for the hard process, Pythia8 (v8.243) for showering, and Delphes (v3.4.2) for detector simulation. The full configurations are also found in the dataset release (Raine et al., 2023).

Two sources of information loss make the inverse problem under-constrained. First, the two neutrinos in dileptonic $t\bar{t}$ decays escape detection entirely and only their summed transverse component is observable as missing transverse momentum $\left(\vec{E}_T^{miss}\right)$, which provides two constraints in place of the eight neutrino degrees of freedom. Second, parton-jet matching is combinatorial: the two reconstructed bottom-jets can each be associated with either the top or antitop, with no detector-level information distinguishing the two assignments.

Detector effects (energy smearing, angular resolution, jet reconstruction efficiency, bottom-tagging etc.) introduce additional measurement noise on top of the structural under-determination.

D.5 MODEL ARCHITECTURES

Architectural and training hyperparameters for all five methods are summarized in table 7 and table 8. This section provides supplementary detail and design rationale where relevant.

Analytic solver We use the algebraic on-shell-mass solver of Sonnenschein (2005) as implemented in anonymous.4open.science/r/analyticsolver-607E (MIT License). The solver returns up to 4 discrete neutrino-momentum solutions per event satisfying the on-shell mass constraints. We select the solution per event with $m_{t\bar{t}}$ closest to $2m_t$. The solver returns no solution for $\sim 10\%$ of events and these events are excluded from all metrics and spectra for fair comparison (appendix D.7).

Transformer regression The transformer regression is inspired by CMS Collaboration (2025a), predicting the top and antitop four-momenta directly from the input tokens. The pure-MSE variant uses standard squared-error loss; the MSE + MMD variant adds two MMD penalties on the $m_{t\bar{t}}$ and $p_T^{t\bar{t}}$ marginals (table 8, "MMD regularization" block). We attempted MMD penalties on c_{hel} and $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ across multiple bandwidth and weight configurations and in all cases the MMD term stalled during training without meaningful decrease while MSE continued to improve normally. Marginal regularization requires moving conditional-mean predictions into vanishing per-event posterior density, fighting the MSE objective rather than complementing it when the problem is multimodal. The two losses are structurally incompatible on multimodal observables.

Table 7: Benchmark II model architectures. All learned models share the same input representation and training data; they differ in encoder backbone, output head, and training objective.

Model	Backbone	Output head	Loss	Posterior
Analytic solver	—	On-shell mass system	—	up to 4 discrete solutions
Transformer (MSE)	Transformer encoder	MLP $\to \hat{\boldsymbol{z}} \in \mathbb{R}^{D_z}$	MSE	$\delta(\hat{\boldsymbol{z}})$
Transformer (MSE+MMD)	Transformer encoder	MLP $\to \hat{\boldsymbol{z}} \in \mathbb{R}^{D_z}$	MSE + λ MMD	$\delta(\hat{\boldsymbol{z}})$
Discrete NF	Transformer encoder	ARQS coupling layers	NLL (exact)	$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$
Continuous NF	Transformer encoder	Conditional vector field	CFM*	$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ (ODE)

* Trained with conditional flow matching (CFM) (Lipman et al., 2023).

Discrete normalizing flow The discrete flow is inspired by the ν^2-flow architecture (Raine et al., 2024): a transformer condition encoder feeds into a stack of conditional ARQS coupling layers with LU-linear permutations. Targets are normalized via an iterative mean-variance layer before flow training, with the corresponding Jacobian included in the likelihood evaluation. The base distribution is standard Gaussian.

Continuous normalizing flow The continuous flow uses the conditional flow-matching objective (Lipman et al., 2023). The vector field $v_\theta(\boldsymbol{z}_t, t, \boldsymbol{x})$ is parameterized as an encoder-decoder transformer. An encoder processes the conditioning tokens and a decoder cross-attends to the encoded condition while conditioning on the integration time via a Gaussian Fourier embedding (table 8). Sampling and likelihood evaluation use ODE integration and details are listed in the same table. Our transformer adapts baseline-transformer building blocks from the L-GATr codebase (Brehmer et al., 2025; Spinner et al., 2024; Brehmer et al., 2023).

D.6 Training details

The models used in Benchmark II are shown in table 7. Reference implementations of all four models, training scripts, and the exact configuration files used to produce table 3 and fig. 2 can be found in the repository referenced in section 8. All models are trained with AdamW (Loshchilov & Hutter, 2017) with hyperparameters shown in table 8.

D.7 Evaluation details

The analytic solver returns a solution for $\sim 90\%$ of events; we restrict marginal-spectrum and metric computation to its solvable subset for fair comparison. The regression and flow-based models reconstruct all events (100% efficiency), and their metrics on the full test set differ by only a small amount. For the flows we draw $N = 500$ posterior samples per event: CRPS uses the full ensemble; χ^2_{spec} uses one random draw per event. Histograms use $B = 100$ bins (only bins with nonzero expected content enter in computation) for calculating χ^2_{spec}. Conformal calibration uses $n_{\text{cal}} = 1{,}000$ events from the test set, with the remaining events for evaluation. We use the per-event NLL as nonconformity score for both flows.

CRPS sensitivity to posterior ensemble size Based on the discussion in appendix C.5 we also include a plot showing dataset-averaged CRPS as a function of the number of posterior samples M used for its estimation in fig. 6. The estimator stabilizes well below the per-method gaps reported in table 3: by $M \approx 100$ all curves are within Monte Carlo noise of their asymptotic values, justifying the choice of $M = 500$ in the main text and confirming that method rankings are robust to ensemble size. The estimator bias from finite N is $\mathcal{O}(1/N) \approx 0.05\%$ (Ferro et al., 2008) when $M = 500$.

Robustness of χ^2_{spec} To assert that the observed ranking on χ^2_{spec} is not an artifact, we test its robustness to binning and sampling conventions.

Bin count: The metric requires a bin count, fixed at $B = 100$ in table 3. We sweep $B \in \{25, \ldots, 200\}$ to verify that the inter-method gaps are not binning-dependent (table 9). Absolute values decrease with B for all models, motivating relative comparisons rather than raw values. Across the range, the flows stay stable while the regression models never approach them.

Table 8: Benchmark II training hyperparameters. Three architectures: a transformer encoder for the MSE/MMD regressions, an encoder-decoder vector field for the continuous flow, and a ν^2-flow-style (Raine et al., 2024) architecture combining a transformer condition encoder with a stack of RQS coupling layers for the discrete flow.

	Regression (MSE & MMD)	Continuous flow (flow matching)	Discrete flow (ν^2-flow style)
Condition encoder			
Encoder blocks	8	4	4
Attention heads	8	8	8
Hidden dimension	128	128	128
Dropout	0.1	0.1	0.1
Positional encoding dim	8	8	8
Flow / decoder head			
Decoder blocks	—	2	—
Time embedding	—	Gaussian Fourier (dim 8, scale 30)	—
Cross-attention	—	decoder $\rightarrow$ encoded condition	—
Coupling layers	—	—	10 RQS stacks
Conditioner MLP	—	—	2 layers, width 128
Spline bins / tail bound	—	—	8 / ± 3.0
Permutations	—	—	LU linear (between stacks)
Initialization	—	Gaussian fitted to data	identity
Base distribution	—	Gaussian	Gaussian
Target normalization	—	—	IterativeNorm (on outputs)
Shared input representation			
Input tokens	6 jets and 2 leptons (4-momenta) + 1 MET (2D transverse vector); per-token 6-dim PID[†] and 1-dim b-tag score when active. Padding handled by attention masks.		
positional encoding	Sinusoidal positional encoding with $d = 8$.		
Training (all learned methods)			
Optimizer	AdamW $(0.9, 0.999)$		
Learning rate	10^{-3} (cosine schedule)		
Batch size	1024		
Early stopping	Patience $P = 15$ on validation loss evaluated every 1500th batch		
Weight decay	10^{-5}		
Hardware	`NVIDIA A100 80GB PCIe`		
Training time per model	~ 4 h	~ 8 h	~ 16 h
MMD regularization (MMD-regularized transformer only)			
Total loss	$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{m_{t\bar{t}}} \cdot \text{MMD}^2_{m_{t\bar{t}}} + \lambda_{p_T} \cdot \text{MMD}^2_{p_T^{t\bar{t}}}$		
Penalty weights	$\lambda_{m_{t\bar{t}}} = 2 \times 10^4$, $\lambda_{p_T} = 2 \times 10^3$		
Kernel	Multi-scale RBF, biased estimator		
Bandwidths $(m_{t\bar{t}})$	$\{2, 20, 100, 200, 400, 2000\}$ GeV		
Bandwidths $\left(p_{\text{T}}^{t\bar{t}}\right)$	$\{0.8, 8, 40, 80, 160, 800\}$ GeV		
Targeted observables	$m_{t\bar{t}}$ and $p_T^{t\bar{t}}$ only; c_{hel} and $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ are not regularized[*]		
Continuous flow			
ODE solver	Runge-Kutta 4, rel/abs tolerance $10^{-7}/10^{-9}$		
Integration steps	50		
Hutchinson's samples	1		
Sampling			
Sampling time per event	—	~ 2.5 ms without NLL computation	~ 0.5 ms

[*] MMD penalties on c_{hel} and $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ were attempted but did not converge (see appendix D.5); [†]PID is a one-hot-style 6D particle ID encoding of charge and particle type.

Binning-free check: In table 10, we report the Earth Mover's distance alongside χ^2_{spec}. The table is ordered from best (top) to worst (bottom) and is almost identical for both the binned (χ^2_{spec}) and unbinned metric (EMD).

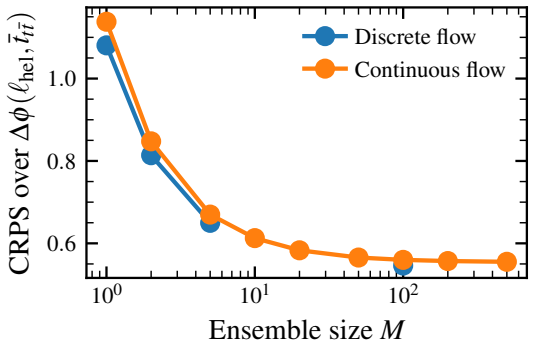


Figure 6: The sensitivity of the CRPS score over $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$ is measured by calculating it as a function of ensemble size, M, for the flow-based models.

Table 9: Bin-count sweep for $\chi^2_{\rm spec}$/ndf on the Benchmark II observable $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$. The values are $\chi^2_{\rm spec}$ while ndf $= B - 1$.

Number of bins B (ndf $= B - 1$)	25	50	75	100	150	200
Analytic solver	10.15	5.96	5.07	4.65	3.62	3.33
Transformer (pure MSE)	605.96	300.29	199.42	150.42	100.08	76.01
Transformer (pure MSE+MMD)	444.31	221.28	146.98	111.20	74.43	56.36
Discrete flow	2.90	2.63	2.38	2.63	2.40	2.43
Continuous flow	2.52	2.26	2.41	2.30	2.06	2.08

Sampling convention: For generative models, the predicted marginal can be constructed by sampling M draws from the per-event posterior. In table 11 we compare sampling $M = 1$ sample per-event with using the full ensemble $M = 500$. Using the full ensemble reduces the flows' χ^2/ndf by averaging Poisson fluctuations. In tables 3 and 12, we use the single-draw convention as the conservative choice. The ranking is consistent across all sampling approaches.

D.8 Results across all four observables

The main text focuses on $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$ as a strongly multimodal observable where our evaluation protocol has the highest impact. This section presents the full evaluation across all four observables ($p_{\rm T}^{t\bar{t}}$, $m_{t\bar{t}}$, $c_{\rm hel}$, $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$, ordered by increasing multimodality of the per-event posterior). Table 12 shows that the ranking flip described in section 6 is consistent across all observables: the pure-MSE transformer wins or near-wins RMSE everywhere disregarding posterior means (see appendix D.8.1), while the flows dominate CRPS and $\chi^2_{\rm spec}$.

The MMD penalty pattern The MMD-regularized transformer reveals the mechanism predicted in section 3.2. On $p_{\rm T}^{t\bar{t}}$, where the per-event posterior is predominantly unimodal, the MMD penalty achieves $\chi^2_{\rm spec}$/ndf, comparable to the flows. On the multimodal $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$, the same penalty produces $\chi^2_{\rm spec}$/ndf in the thousands. This asymmetry is the empirical signature of the law of total variance: marginal regularization can compensate global distortion when within-posterior variance is small, but cannot correct per-event multimodality when within-posterior variance is large. The intermediate observables ($m_{t\bar{t}}$, $c_{\rm hel}$) show intermediate behavior, with MMD reducing but not closing the gap to the flows.

Marginal spectra Figure 7 visualizes the same conclusion. The MSE transformer compresses the marginal of every observable, with the distortion most visually severe for $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$. The MMD transformer recovers the unimodal $p_{\rm T}^{t\bar{t}}$ and $m_{t\bar{t}}$ marginals, where it was specifically tuned (appendix D.5), deviates noticeably for $c_{\rm hel}$, and fails on $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$. Both flows track all four marginals closely, with no observable-specific tuning.

Table 10: Comparison between the binned χ^2_{spec} and unbinned Earth Mover's Distance (EMD) on the Benchmark II observable $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$. The values are mean $\pm$ standard deviation across 4 random seeds and identical to the values presented in table 3. The rows are ordered by χ^2_{spec} from best (top) to worst (bottom).

	χ^2_{spec}/ndf (closest-to-1)	EMD $\downarrow$
Continuous flow	2.550 ± 0.295	0.015 ± 0.002
Discrete flow	2.687 ± 0.447	0.013 ± 0.006
Discrete flow (post. median)	72.726 ± 4.560	0.265 ± 0.009
Continuous flow (post. median)	75.300 ± 0.127	0.270 ± 0.001
Transformer (MSE+MMD)	118.129 ± 4.962	0.330 ± 0.008
Transformer (pure MSE)	142.281 ± 7.680	0.362 ± 0.008
Discrete flow (post. mean)	291.863 ± 6.157	0.539 ± 0.006
Continuous flow (post. mean)	309.315 ± 5.069	0.555 ± 0.004

Table 11: Effect of the sampling convention on χ^2_{spec}/ndf on the Benchmark II observable $\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$ at $B = 100$. The single-draw column reports mean and standard deviation over 25 repetitions.

	One draw per event	500 draws ($1/500$ weighted)
Discrete flow	2.658 ± 0.314	1.693
Continuous flow	2.479 ± 0.335	1.494

Variance compression of spectra The variance compression of section 3.2 is visually clearest on $p_{\text{T}}^{t\bar{t}}$ in fig. 7(a): the MSE transformer's marginal is sharper at the peak and lower in the tail than the truth, with the peak itself shifted upward because the right-tailed underlying distribution pulls conditional means toward higher $p_{\text{T}}^{t\bar{t}}$ values. The same pattern is also more pronounced in $m_{t\bar{t}}$ in fig. 7(b).

Caveat on χ^2_{spec} for heavy-tailed observables The χ^2_{spec} values for $m_{t\bar{t}}$ and $p_{\text{T}}^{t\bar{t}}$ reported in table 12 are systematically biased. Both observables have heavy tails with bins where the Poisson approximation $\sigma_b^2 \approx n_b^{\text{true}}$ underestimates the true uncertainty, inflating χ^2. A proper goodness-of-fit assessment for tail-heavy spectra uses the Baker-Cousins likelihood-ratio statistic (Baker & Cousins, 1984), which we leave for future work. Within-method comparisons remain valid because the bias affects all methods equally on a given observable.

D.8.1 Jensen gap between latent and observable estimators

The conditional-mean pathology of section 3.1 has a basis-dependent refinement when downstream quantities are nonlinear functions of the latent. Let $\mathcal{O} : \mathbb{R}^{D_z} \to \mathbb{R}$ be a physics observable computed from the latent four-vector $\boldsymbol{z}$. The MSE-optimal estimator for $\mathcal{O}(\boldsymbol{z})$ given observations $\boldsymbol{x}$ is

$$\hat{\mathcal{O}}^*(\boldsymbol{x}) = \mathbb{E}[\mathcal{O}(\boldsymbol{z}) \mid \boldsymbol{x}] = \int \mathcal{O}(\boldsymbol{z})\, p(\boldsymbol{z} \mid \boldsymbol{x})\, d\boldsymbol{z}. \tag{16}$$

A regression trained on the latent under squared-error loss instead converges to $\hat{\boldsymbol{z}}(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{z} \mid \boldsymbol{x}]$ and yields the prediction $\mathcal{O}(\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{x}])$ for the observable. By Jensen's inequality,

$$\mathcal{O}(\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{x}]) \neq \mathbb{E}[\mathcal{O}(\boldsymbol{z}) \mid \boldsymbol{x}] \tag{17}$$

in general, with the gap vanishing only when $\mathcal{O}$ is linear or the posterior $p(\boldsymbol{z} \mid \boldsymbol{x})$ is degenerate. The regression is therefore not MSE-optimal for the observable even setting aside the multimodality issues of section 3.2.

Generative methods, by contrast, recover the MSE-optimal observable estimator at evaluation time: given posterior samples $\boldsymbol{z}^{(k)} \sim p(\boldsymbol{z} \mid \boldsymbol{x})$, the Monte Carlo average $\frac{1}{N}\sum_k \mathcal{O}(\boldsymbol{z}^{(k)})$ converges to $\mathbb{E}[\mathcal{O}(\boldsymbol{z}) \mid \boldsymbol{x}]$. This explains the systematic RMSE advantage of the "posterior mean" rows in table 12. Across all four observables, both flow posterior means achieve RMSE substantially below all latent-space regression baselines, despite the regressions being trained directly to minimize four-vector

Table 12: Benchmark II: methods evaluated under pointwise (RMSE) and distributional (CRPS, χ^2_{spec}) metrics across four observables (ordered by increasing per-event posterior multimodality). Values are mean $\pm$ standard deviation across 4 random seeds. Averaging the flow posterior *in observable space* per event recovers RMSE substantially better than any latent-space regression, demonstrating the Jensen-gap argument of appendix D.8.1: the MSE-optimal observable estimator is the conditional mean of the observable, not the observable of the conditional mean. The χ^2 for $p_{\text{T}}^{t\bar{t}}$ and $m_{t\bar{t}}$ is biased since the Gaussian approximation breaks down in the heavy tails (Baker-Cousins (Baker & Cousins, 1984) would be more appropriate, but within-method comparisons remain valid). Best per column in **bold**

Method	RMSE↓	CRPS↓	χ^2_{spec}/ndf↓
	$p_{\text{T}}^{t\bar{t}}$		
Analytic solver	34.20	20.97	1,450 / 82
Transformer (pure MSE)	36.85 ± 0.24	25.31 ± 0.19	$6{,}451 \pm 546$ / 82
Transformer (MSE+MMD)	35.06 ± 0.27	24.41 ± 0.24	663 ± 180 / 82
Discrete flow	59.30 ± 41.62	$\mathbf{12.35} \pm 0.22$	$\mathbf{650} \pm 69$ / 82
Discrete flow (post. mean)	$\mathbf{25.96} \pm 0.47$	17.39 ± 0.57	$10{,}016 \pm 944$ / 82
Continuous flow	42.07 ± 0.92	13.17 ± 0.14	985 ± 85 / 82
Continuous flow (post. mean)	27.22 ± 0.29	19.50 ± 0.25	$15{,}671 \pm 1{,}348$ / 82
	$m_{t\bar{t}}$		
Analytic solver	116.57	71.55	4,183 / 65
Transformer (pure MSE)	93.03 ± 0.50	66.47 ± 0.99	$16{,}923 \pm 2{,}378$ / 68
Transformer (MSE+MMD)	94.45 ± 0.52	64.26 ± 0.59	$1{,}879 \pm 1{,}248$ / 68
Discrete flow	$363.72 \pm 445.68^{\ddagger}$	38.28 ± 0.36	767 ± 231 / 68
Discrete flow (post. mean)	149.09 ± 134.19	59.92 ± 14.44	$10{,}020 \pm 580$ / 68
Continuous flow	119.45 ± 6.79	$\mathbf{37.70} \pm 0.15$	$\mathbf{586} \pm 245$ / 68
Continuous flow (post. mean)	$\mathbf{81.02} \pm 0.37$	52.69 ± 0.31	$7{,}964 \pm 414$ / 68
	c_{hel}		
Analytic solver	0.412	0.303	416 / 99
Transformer (pure MSE)	0.402 ± 0.001	0.299 ± 0.001	$2{,}457 \pm 181$ / 99
Transformer (MSE+MMD)	0.414 ± 0.003	0.311 ± 0.003	$3{,}457 \pm 219$ / 99
Discrete flow	0.448 ± 0.002	$\mathbf{0.166} \pm 0.001$	$\mathbf{189} \pm 21$ / 99
Discrete flow (post. mean)	$\mathbf{0.319} \pm 0.001$	0.241 ± 0.001	$10{,}526 \pm 257$ / 99
Continuous flow	0.459 ± 0.007	$\mathbf{0.166} \pm 0.001$	247 ± 21 / 99
Continuous flow (post. mean)	$\mathbf{0.319} \pm 0.001$	0.244 ± 0.001	$12{,}352 \pm 1{,}180$ / 99
	$\Delta\phi(\ell_{\text{hel}}, \bar{t}_{t\bar{t}})$		
Analytic solver	1.740	1.036	462 / 99
Transformer (pure MSE)	1.499 ± 0.005	0.941 ± 0.003	$14{,}086 \pm 760$ / 99
Transformer (MSE+MMD)	1.500 ± 0.011	0.936 ± 0.006	$11{,}695 \pm 491$ / 99
Discrete flow	1.766 ± 0.005	$\mathbf{0.544} \pm 0.001$	266 ± 44 / 99
Discrete flow (post. mean)	$\mathbf{1.255} \pm 0.002$	0.854 ± 0.002	$28{,}894 \pm 610$ / 99
Continuous flow	1.780 ± 0.006	0.546 ± 0.001	$\mathbf{252} \pm 29$ / 99
Continuous flow (post. mean)	$\mathbf{1.255} \pm 0.001$	0.866 ± 0.001	$30{,}622 \pm 502$ / 99

‡RMSE on $m_{t\bar{t}}$ is dominated by far tail posterior draws which the squared-error statistic is highly sensitive to. In contrast, the distributional metrics on the same row are stable to within a few percent.

MSE. The Jensen gap for these observables is large because top-quark posteriors are wide and the observable maps are strongly nonlinear (squared invariant masses, transverse-plane angles).

This argument is independent of the multimodality argument of section 3.2: even for unimodal posteriors with substantial variance, the Jensen gap persists. It provides a second, independent reason to prefer methods that preserve the per-event posterior over those that collapse it to a four-vector point estimate. The argument also highlights why the basis choice for any latent state regression directly influences the reconstruction RMSE on observables.

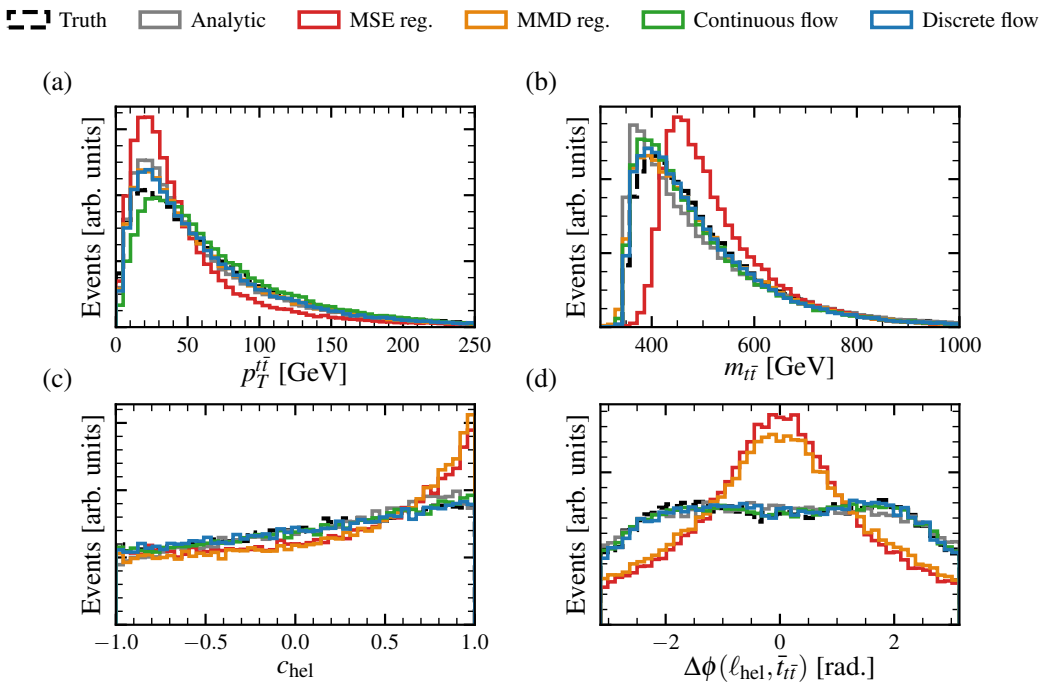


Figure 7: Marginal reconstructed spectra for $p_{\mathrm{T}}^{t\bar{t}}$, $m_{t\bar{t}}$, c_{hel}, and $\Delta\phi(\ell_{\mathrm{hel}}, \bar{t}_{t\bar{t}})$ (in order of increasing multimodality). The dashed line shows the truth marginal and the colored curves show each method's reconstructed marginal. Quantitative comparison via χ^2_{spec} is in table 12.

D.9 Per-event posterior structure across observables

Figure 8 extends fig. 2(a) by showing per-event reconstructed posteriors for three randomly selected test events across all four observables. The same per-event collapse pattern observed for $\Delta\phi(\ell_{\mathrm{hel}}, \bar{t}_{t\bar{t}})$ in the main text persists across all observables: both transformer regressions concentrate near the per-event posterior mean, while the flows recover the underlying multimodal or asymmetric structure where present. The contrast is most visually striking on $\Delta\phi(\ell_{\mathrm{hel}}, \bar{t}_{t\bar{t}})$ (strongly multimodal). On $p_{\mathrm{T}}^{t\bar{t}}$ and $m_{t\bar{t}}$ the posteriors are predominantly unimodal but skewed, and the regression's bias is correspondingly milder. Event 135 (bottom-right in the $\Delta\phi(\ell_{\mathrm{hel}}, \bar{t}_{t\bar{t}})$ row) is the same event shown in fig. 2(a). Across all events, the analytic solver returns at most a single point estimate per event and fails to return any solution for Event 0.

D.10 Conditional calibration reveals architecture and data limitations

Section 4.3 recommends assessing calibration conditionally – in bins of phase space – to identify regions where uncertainty estimates fail despite acceptable global coverage. Figure 9 applies this diagnostic to the two flows, plotting empirical coverage at the 90% nominal level as a function of each observable. Both flows undercover throughout, but the discrete flow tracks the nominal level substantially more closely than the continuous flow across all four observables, consistent with the marginal coverage curves of fig. 2(c). The continuous flow's undercoverage is approximately uniform in observable value, suggesting a global density-evaluation issue (likely from approximate ODE-based likelihood; see section 6) rather than a phase-space-specific failure. The exception is the high-$p_{\mathrm{T}}^{t\bar{t}}$ region ($p_{\mathrm{T}}^{t\bar{t}} \gtrsim 450\,\mathrm{GeV}$), where both flows undercover more severely. This regime is sparsely populated in the training data, and the degraded calibration likely reflects extrapolation rather than a model failure. Conditional calibration thus surfaces both a global architectural difference (continuous vs discrete) and a phase-space-specific data limitation, neither of which is visible in the global coverage curve alone.

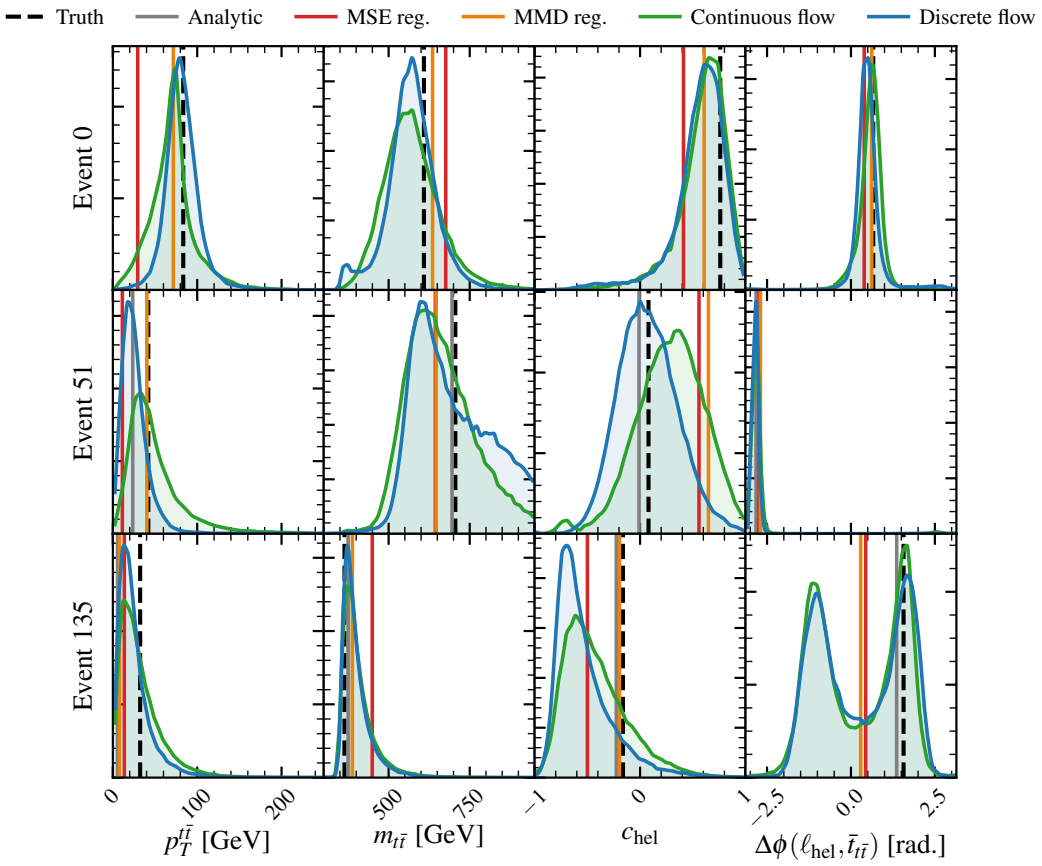


Figure 8: Reconstructed posteriors (filled curves) and point estimates (vertical lines) for three randomly selected test events across the four observables $p_{\mathrm{T}}^{t\bar{t}}$, $m_{t\bar{t}}$, c_{hel}, and $\Delta\phi(\ell_{\mathrm{hel}}, \bar{t}_{t\bar{t}})$. True targets are shown as dashed black lines. Event 135 (bottom row of $\Delta\phi(\ell_{\mathrm{hel}}, \bar{t}_{t\bar{t}})$ column) is the event displayed in fig. 2(a).

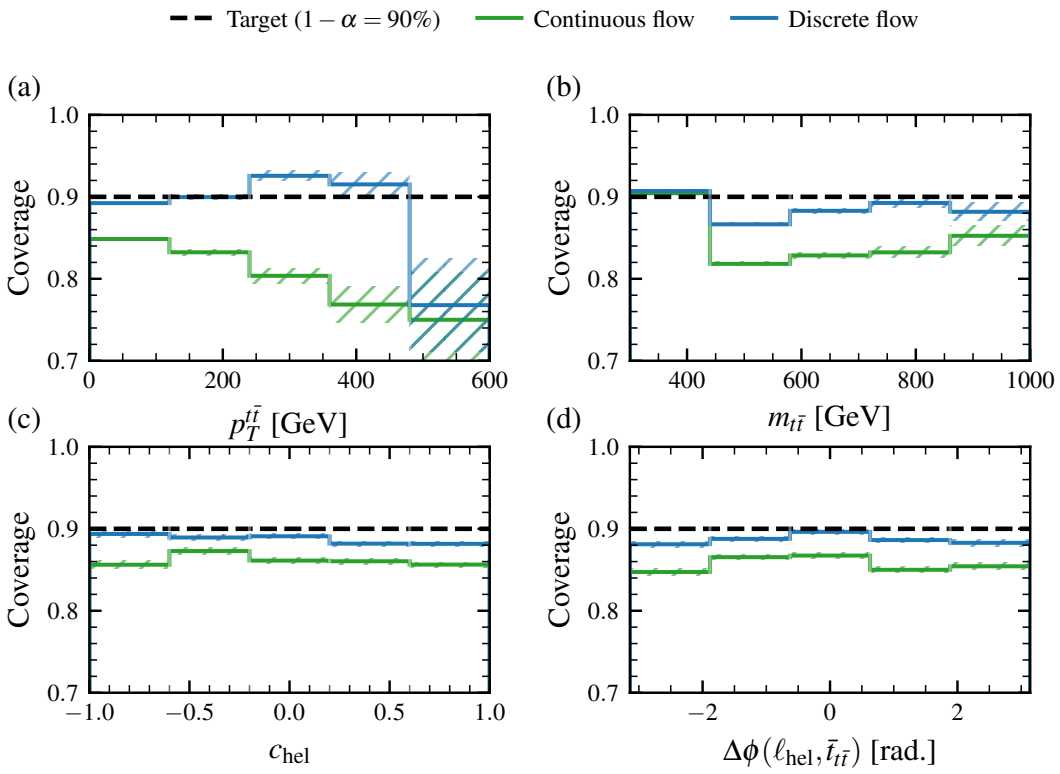


Figure 9: Empirical coverage at the 90% nominal level as a function of the true value of each observable. Shaded bands show binomial uncertainty per bin. The rightmost bins for $p_{\text{T}}^{t\bar{t}}$ and $m_{t\bar{t}}$ have low statistics as indicated by the uncertainty bands.

Table 13: Joint evaluation of the 8-dimensional latent of Benchmark II on a single seed. The energy score (ES) is reported in two normalizations: raw four-momenta and whitened by component standard deviation. The RMSE (mean) column applies only to the flows where the posterior mean is scored. Best per column in **bold**.

	RMSE 1-samp [GeV]↓	RMSE mean [GeV]↓	ES [GeV]↓	ES [σ]↓
Analytic solver	133.161	—	267.715	1.355
Transformer (pure MSE)	**109.373**	—	235.218	1.169
Transformer (MSE+MMD)	109.962	—	237.736	1.180
Continuous flow	166.380	106.750	160.982	0.791
Discrete flow	214.708	107.130	**156.665**	**0.764**

D.11 Joint evaluation of the 8-dimensional latent

Section 6 scores the univariate observable $\Delta\phi(\ell_{\rm hel}, \bar{t}_{t\bar{t}})$. Axis I extends to the joint latent through the energy score (Gneiting et al., 2008) – the multivariate generalization of CRPS – which is strictly proper and reduces to the Euclidean norm of the error for point estimators preserving cross-family comparability. Here, we evaluate it on the full latent $z \in \mathbb{R}^8$ using 500 posterior samples per event.

In contrast to the one-dimensional case, the energy score also depends on how the individual components are normalized. We report two normalizations here: raw four-momentum components in GeV, and components whitened by the truth standard deviation, which places differing scale components on a common footing. Alongside, we report the 8-dimensional RMSE. The energy score results are shown in table 13.

The model ranking shown in table 3 is preserved in the full latent space. The regression models win on RMSE while the flow-based models win on energy score. The ranking flip is therefore not a result of the one-dimensional scoring of a derived observable. In addition, averaging the flow posteriors improves RMSE over the regression models which is the Jensen Gap effect (appendix D.8.1) also observed for the one-dimensional scoring.

This instantiates Axis I in a joint setting, but does not extend Axis II. A population fidelity metric in all eight dimensions requires an unbinned distance such as EMD or MMD on the joint marginal which we do not validate here (section 7).

E Reproducibility

This appendix documents the artifacts released with the paper. Beyond reproducing the reported experiments, the released evaluation pipeline and protocol documentation are intended to let practitioners apply the three-axis protocol to their own supervised reconstruction models, on benchmarks beyond the two studied here.

E.1 Code repository and structure

The code accompanying this paper is released at anonymous.4open.science/r/evaluating-generative-models-C559/ (MIT License).[2] The repository contains the implementations of all learned models in Benchmark II, the training and evaluation pipelines, the reference implementations of the three protocol metrics, and the analysis notebooks used to produce every table and figure in sections 5 and 6. Compute requirements, training times, and hyperparameters are summarised in appendices C.3 and D.6. Executed notebooks and the pre-computed evaluation arrays underlying every Benchmark II table and figure are archived on Zenodo[3] under CC-BY 4.0, enabling readers to regenerate the plots and verify reported metrics without rerunning training or posterior sampling: `https://zenodo.org/records/22963037?token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6ImRlYmRmMjg0LWMyM2ItNGRiNS05MzgwLWE0MzBiM2NkY2RkNCIsImRhdGEiOnt9LCJyYW5kb20iOiIzMzZkZmQyNTc3Y2I1ODQwYmJjODFkNmJlM2JiNDIzMCJ9.jESGrjRGBOlQ4LQL`

[2] For double-blind review, hosted via anonymous.4open.science; the public URL will be substituted on acceptance.

[3] For double-blind review, we provide a tokenized Zenodo link. DOI will be substituted on acceptance.

zQoKeY6GVBmEFX89SRsc1T2D5l7UmoMkkKFv8VoMD5rQzDBYdiFMdDQCQsG9mEKSV0zzMg

Repository layout

- `README.md` A more practical oriented description on how the repository is set up, how to install the virtual environment, and how to run different parts of the code base.
- `src/top_reco/` – the core Python package for the studies presented in Benchmark II. The `models/` subpackage contains the model implementations; `losses/`, `tasks/` and `data/` provide the training objectives, the dataset wiring and the Delphes-format data loader.

  The `metrics/` subpackage provides reference implementations of CRPS (section 4.1), spectrum fidelity (section 4.2), and the conformal calibration diagnostic (section 4.3), each independently testable and importable as `from top_reco.metrics import ...`.
- `configs/` – The model configuration tree with hyperparameters for each model.
- `scripts/` – `train.py` (entry point for training a model), `evaluate.py` (produces the per-event metric arrays), `posterior_sampling.py` (produces the dense posterior samples used for the per-event posterior figures).
- `notebooks/` – `benchmark1_synthetic.ipynb` is end-to-end and self-contained (no external data). It produces all tables and figures related to Benchmark I. `benchmark2_plotting.ipynb` reads pre-computed evaluation arrays from `data/benchmark2/` and produces all tables and figures related to Benchmark II.
- `tests/` – unit tests for the `coordinates/` and `metrics/` modules.
- `docs/` – Sphinx-built API reference for the metrics module, generated from the same docstrings that document the formulae of section 4.
- `NOTICE.md` and `LICENSES/` – third-party code attribution; portions of `src/top_reco/coordinates/` and the transformer building blocks of `src/top_reco/models/` are adapted from other sources.

The *analytic solver* baseline used in Benchmark II is the algebraic on-shell-mass solver of Sonnenschein (2005); the implementation is external to this repository and is available at anonymous.4open.science/r/analyticsolver-607E (MIT License).[4]

E.2 Compute resources

All experiments were run on a single `NVIDIA A100 80GB PCIe` GPU. Per-model training times are: Benchmark I – MLP-backbone models 1 minute each, normalizing flow 7 minutes (5 seeds each); Benchmark II – transformer regressions 4 hours, continuous flow 16 hours, discrete flow 8 hours (4 seeds each, table 8). In addition, the flow-based models needed GPU-time for sampling which we estimate to a combined sampling time of 30 GPU-hours. Total reported-experiment compute is approximately 100 GPU-hours. Total project compute including hyperparameter exploration, MMD-bandwidth scans on additional observables (appendix D.5), and discarded architectural variants is approximately 450 GPU-hours.

E.3 Dependencies and environment

Models are implemented in PyTorch ($\geq$`2.2`). The synthetic-benchmark normalizing flow uses normflows (Stimper et al., 2023). The continuous flow uses standard PyTorch ODE integration utilities (Chen, 2018). Full dependency lists with pinned versions are provided in the repository in `uv.lock`.

[4]For double-blind review, hosted via anonymous.4open.science; the public URL will be substituted on acceptance.

E.4 Dataset licensing and access

The Benchmark II dataset is the public Delphes Monte Carlo release of Raine et al. (2023), distributed under CC-BY 4.0. We use the train/test splits provided with the release without modification, and we restrict ourselves to events generated with MadGraph. The Benchmark I synthetic dataset is generated by the released code from the forward model of eq. (6); the data-generation script is included in the repository and produces deterministic outputs given the documented seeds.