Trusting Generative Unfolding

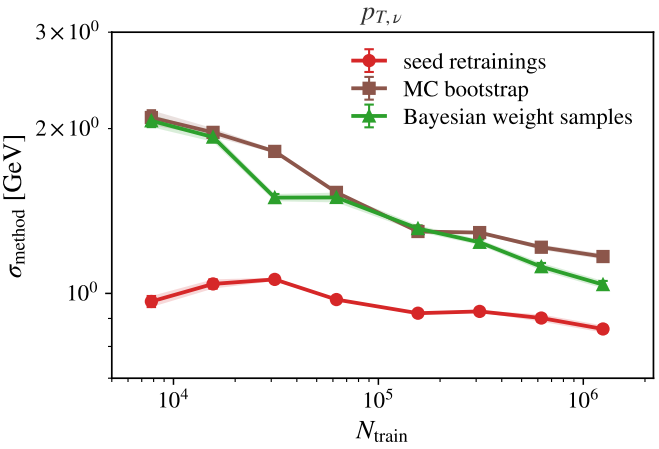
What the uncertainty band of a generative unfolder is made of

Sebastian Pitz
with Anja Butter and Bogdan Malaescu

LPNHE, Sorbonne Université, Université Paris Cité, CNRS/IN2P3, Paris

17 September 2026

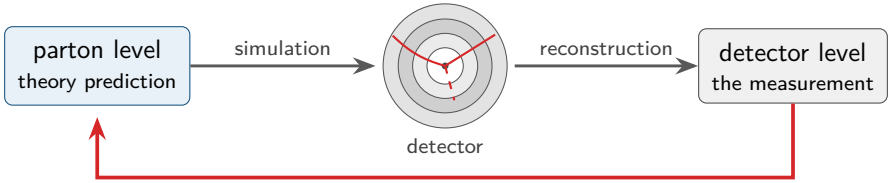
Three uncertainty estimates for one measurement



- One trained unfolder, one training set
- Three published recipes for its error bar
- They differ by up to a factor two

Which estimate is correct?

Unfolding: detector-independent results and their uncertainty



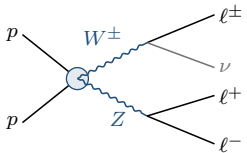
Why unfold

- A measurement is the detector's image of the physics; unfolding inverts the detector once
- The result is compared with any prediction, now or in ten years, without the detector model

What makes it hard

- The inverse is not unique: many parton events give the same record, so the answer is a distribution per event
- The inverse is learned by a network: its training (initialisation, optimisation) and finite sample add uncertainty
- The simulation it learns from is not the data it unfolds (perspective)

The physics case: WZ production and the unfolding problem



Reco —unfold→ **parton level**

20 detector columns — the condition x

9 parton coordinates — the target t

$pp \to W^{\pm}Z \to \ell^{\pm}\nu\,\ell^{+}\ell^{-}$

- Detector record x: lepton momenta, missing transverse momentum
- Parton-level event t: lepton and neutrino four-momenta
- The inverse is a conditional density: one distribution per event, with the simulation as its prior

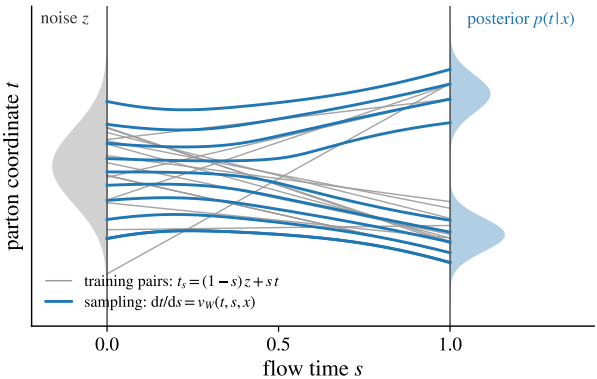
Learned detector inverse, trained on simulated pairs:

$$p_{\text{model}}(t \mid x) \approx p(t \mid x) = p(x \mid t)\, p_{\text{sim}}(t) \,/\, p_{\text{sim}}(x)$$

Unfolded distribution of the measured events:

$$p_{\text{unfold}}(t) = \int \mathrm{d}x\; p_{\text{model}}(t \mid x)\; p_{\text{data}}(x)$$

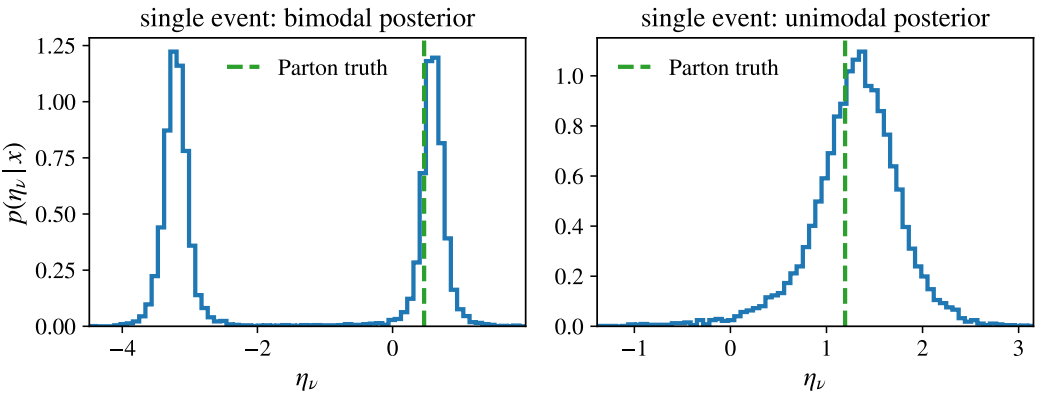
Conditional flow matching: from noise to parton-level events



- Training: regress $v_W(t_s, s, x)$ onto the pair direction $t - z$, conditioned on the detector record x
- Sampling: integrate $\mathrm{d}t/\mathrm{d}s = v_W$ from a noise draw z
- K solutions per record: K draws of the posterior $p(t \mid x)$

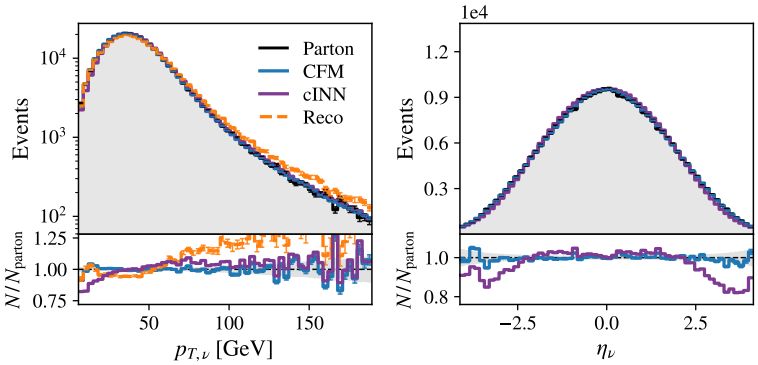
$$\mathcal{L}_{\mathrm{CFM}}(W) = \mathbb{E}_{s,\,z,\,(t,x)} \left\| v_W(t_s, s, x) - (t - z) \right\|^2, \qquad t_s = (1 - s)\,z + s\,t$$

Single-event posteriors: bimodal and unimodal cases



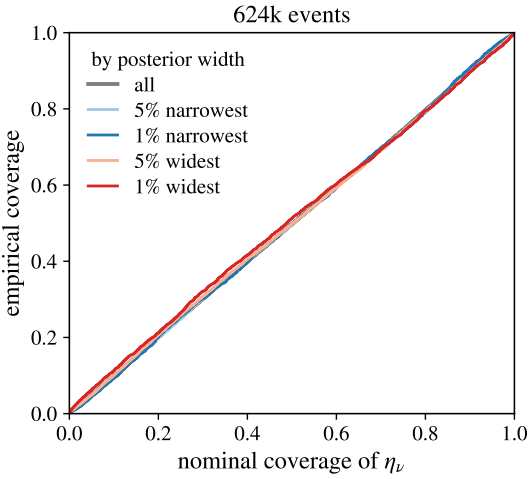
Two single events of the held-out test sample, unfolded by one CFM (10^4 posterior draws each); dashed: the parton-level truth.

Sample-level closure: the CFM inverts the detector response



- Closure: the unfolded spectrum (pooled posteriors, 312k test events) against the parton level; the detector-level input for scale
- Neutrino pseudorapidity: recovered although the detector does not measure it
- cINN baseline at the same data, optimiser steps and parameter count: the architecture, not the budget, sets the tails
- Bands: Poisson (Parton, Reco), test-sample bootstrap (CFM)

Per-event calibration of the posterior width



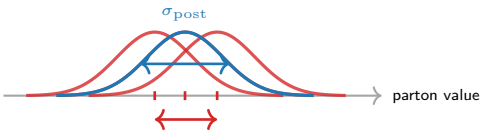
- Question: the reported width as an error bar event by event, not only on average
- Test: rank of the truth among the posterior draws (PIT), in slices of the model's own reported width; the selection never uses the truth
- Result at the production volume: the reported width is calibrated in every η_ν slice, including the 1% of events the model is least certain about; $p_{T,\nu}$ within four points
- Consequence: the reported width is the per-event uncertainty of the unfolder; the spread across trainings is the correction on top

Training-axis uncertainty I: seed retraining and the MC bootstrap

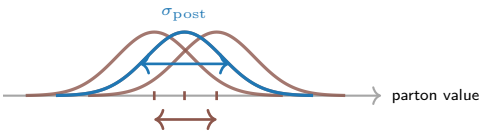
$$\sigma^2_{\text{tot}} = \underbrace{\mathbb{E}_W\left[\text{Var}_z\, T\right]}_{\sigma^2_{\text{post}}} + \underbrace{\text{Var}_W\left[\mathbb{E}_z\, T\right]}_{\sigma^2_{\text{train}}}$$

σ^2_{post}: width of one training's answer

σ^2_{train}: spread of the answer across trainings



σ_{seed}: new initialisation, same sample

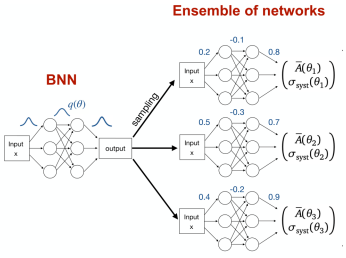


σ_{boot}: Poisson weights, initialisation fixed

- **Seed retraining**: the initialisation and batch order change, the training sample does not
- Measures how far the optimiser's freedom alone moves the answer
- B trainings; the spread stops shrinking above a few thousand events

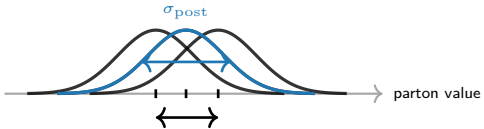
- **MC bootstrap**: a Poisson(1) weight per training event in the loss, $\mathcal{L} = \mathbb{E}\left[\, w_i \, \| v_W - (t - z) \|^2 \right]$; initialisation fixed
- Measures the sensitivity of the answer to the finite training sample
- B trainings, no new simulation

Training-axis uncertainty II: the Bayesian CFM and the new-data reference



sketch: Plehn et al., arXiv:2211.01421, Fig. 5 (after arXiv:2003.11099)

- **Bayesian CFM**: a Gaussian per weight, trained by the expected loss under weight noise plus a prior term,
  $\mathcal{L}_{\text{B}} = \langle \mathcal{L}_{\text{CFM}} \rangle_{w \sim q} + \frac{C}{N_{\text{train}}} \text{KL}(q \,\|\, p)$
- Each member is one draw of the weights: σ_W from a single training, the lowest cost



- **New data**: disjoint, independently simulated training sets of equal size, one training each, initialisation fixed
- Reference: the spread of independently simulated training sets, no resampling, no approximation; σ_{boot} and σ_W must reproduce it
- Costly in data and trainings; done here from 1k to 624k events

Two terms on the training axis and three estimators of the sample term

$$\sigma^2_{\text{tot}} = \underbrace{\mathbb{E}_W\left[\text{Var}_z\, T\right]}_{\text{posterior width}} + \underbrace{\sigma^2_{\text{seed}}}_{\text{seed}} + \underbrace{\sigma^2_{\text{sample}}}_{\text{training sample}}$$

seed: the optimiser

new initialisation, same data; measured by retraining

$\oplus$

training sample: target = new data, initialisation fixed

bootstrap

resample the sample, retrain

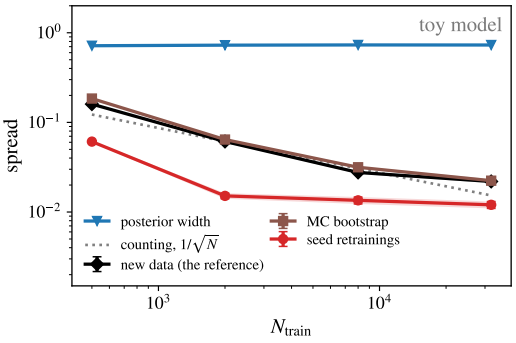
or

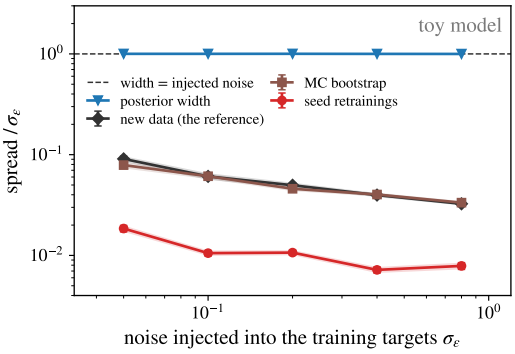
Bayesian

sample the weights of one training

Seed and sample terms add in quadrature; bootstrap or Bayesian estimate the sample term, never both.

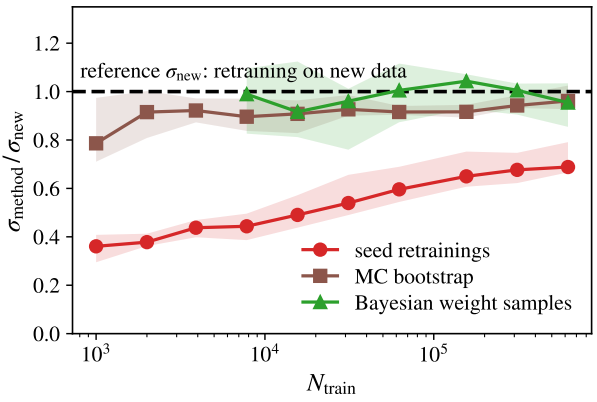
Toy model: training volume and target noise act on separate axes





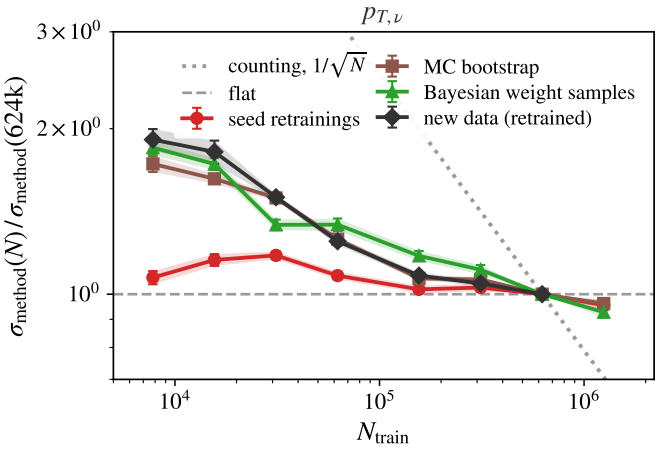
- Toy: a one-dimensional inverse problem with a two-branch posterior (the η_ν analogue), a mixture-density network; new data is free and the true posterior is known
- Training volume moves the training spreads, not the posterior width; the seed spread is an optimiser floor, not a statistical term
- Target noise moves the posterior width one to one; the training spreads are nearly insensitive to it

Do the estimators reproduce the new-data reference?



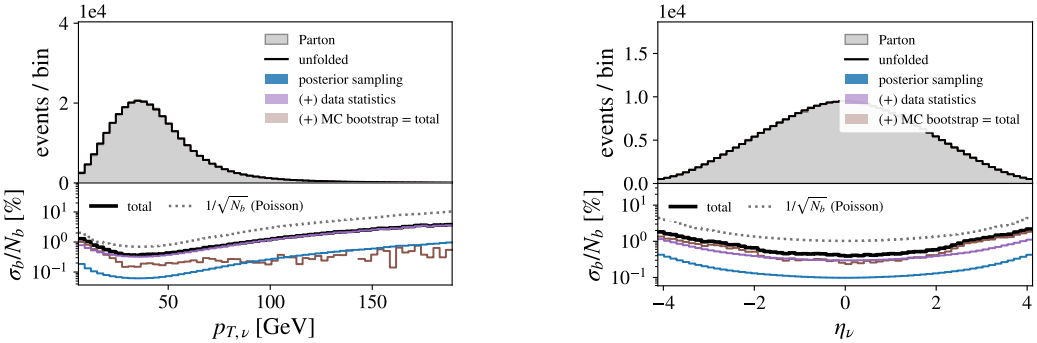
- Test: $\sigma_{\mathrm{method}}/\sigma_{\mathrm{new}}$ at every volume; σ_{new} = spread of retraining on new data; 1 = the estimator reproduces new data; marker: median over seven observables, band: their range
- One procedure for every family, its length chosen once at the reference's optimum
- MC bootstrap: median within 10% of σ_{new} from 2k, within 5% at the production volume; Bayesian band: at the level of σ_{new} at the production volume, at its own budget
- Seed retraining: the other term, added in quadrature; a third to three quarters of σ_{new}

Dependence on the training volume: the two axes separate



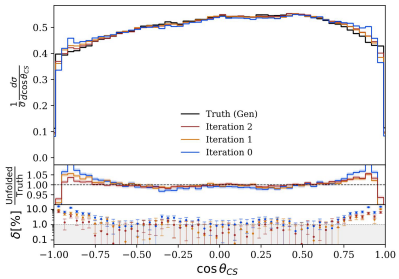
More simulation shrinks the sample term, not the seed term; neither falls as fast as counting.

The training term on the unfolded spectrum



- Per bin b with N_b events, in quadrature: posterior sampling, statistics of the measured sample, training sample (the term verified against new data); production volume
- The verified term dominates the observable the detector does not measure
- Statistical axes only: the unfolder's residual against the parton level is a separate entry

Outlook



Iterate the simulation prior away. Earlier work, SM-trained unfolder on a SMEFT-modified sample: two iterations reach the truth. To do: the error bar round the loop.

Detector systematics

- Vary the missing-momentum scale or resolution
- Propagate it through the fixed unfolder (a shift) or train with it (the posterior width absorbs it, as in the noise toy)

Every new estimator is measured against retraining on new data.

Conclusions

- **The unfolder**: one distribution per event, closing on the parton level, with a width that stays calibrated where it is widest
- **Two axes, one identity**: the posterior width is physics and carries the coverage; the training spread is a percent-level correction on top
- **The cheap estimators**: the bootstrap reproduces retraining on new data; the Bayesian band sits at the same level at the production volume; the seed term is separate and added in quadrature
- **A slow lever**: the training spreads fall at a quarter to a third of the counting rate, so more simulation buys the error bar down only slowly

The error bar of a generative unfolder can be measured, not only estimated.

Thanks for your attention!

Backup

Backup: the three estimators, in numbers

- Hook figure: pooled rms spread of the per-event posterior means over the 312k test events, $p_{T,\nu}$, pedestal corrected; leave-one-member-out (jackknife) errors 1–3%. Members: 8 at 7.8k, 15.6k, 624k and 1.25M, 50 at 31k–312k; Bayesian: 50 weight samples of one matched-step training per volume, drawn from 7.8k where its checkpoints are annealed.
- Seed and bootstrap are ensembles of ordinary point-estimate networks of the production architecture; the Bayesian curve is the same architecture with a Gaussian per weight.
- The page starts at 7.8k: below, the fixed-epoch recipe is 150–300 optimiser steps and the matched-step Bayesian checkpoints are unannealed; the ladder from 1k is on the scaling slide.
- Seed curve: $+10\%$ from 7.8k to 31k, -20% from 31k to 1.25M, significant against the errors. With 150 epochs the seed members are not at their validation optimum between 7.8k and 31k (300-epoch pilots: optimum at epoch 241 and 292 at 15.6k and 31k; bootstrap members at 151, range 121 to 185), so that part of the shape is provisional until the schedule scan.
- Bayesian step at 31k ($1.93 \to 1.50$ GeV): at that cell the band is 0.82 of the bootstrap on $p_{T,\nu}$ and $\eta_{\ell W}$ only (0.97–1.08 on the other five), a second independent Bayesian training at 31k reproduces it to 1%, bootstrap and new data are normal there; a property of the Bayesian recipe at that volume on two observables, mechanism not understood (controls exist at 7.8k only).
- Related work: deep ensembles and bootstrap ensembles are the standard recipes; Bayesian flows are checked against a toy likelihood (Butter, Diefenbacher, Plehn, Vogel 2025) or for histogram coverage across five independent toy training sets, generation only (Bieringer, Diefenbacher, Kasieczka, Trabs 2024); neural posterior unfolding (Torales Acosta et al. 2025) is binned, posterior on the bin counts, no network training uncertainty. None compares the three recipes against retraining on new data for an unfolder.

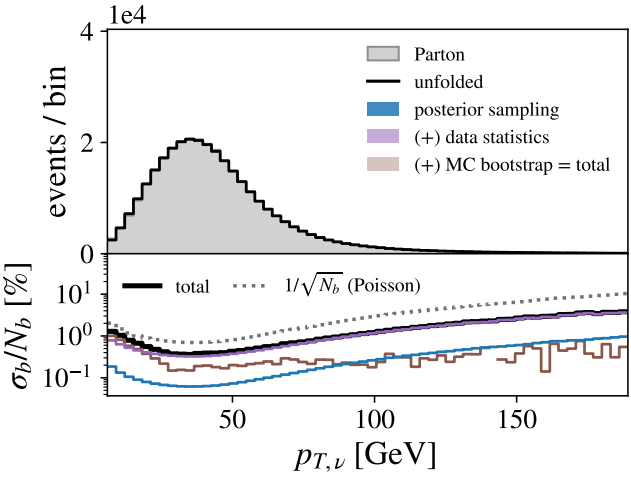
Backup: model, training, coordinate recipes

target	9 parton coordinates
condition	20 detector columns
network	residual MLP, 8 blocks $\times$ width 512
parameters	2.1×10^6 (4.3×10^6 Bayesian)
training	AdamW, batch 4096, lr 4×10^{-4}
schedule	150 epochs, cosine anneal, best-validation restore
data	1,040,953 events, split 60/10/30
	$N_{\text{train}} = 624{,}000$, $D = 312{,}286$ test
sampling	`dopri5` ODE, $K_z = 100$ latent samples per event

Bayesian variant $\langle \mathcal{L}_{\text{CFM}} \rangle_{W \sim q} + (C/N_{\text{train}})\,\mathrm{KL}(q\|p)$,
$C = 0.1$ tuned at 624k; 50 weight samples $\times$ 30 latent draws.

- **WRAP** (default): azimuths rotated so the leading Z lepton sits at $\phi = 0$ and offset by π so the periodic boundary falls in a density trough; the network predicts the residual to the matched detector value, periodic residuals wrapped onto one period.
- **ZC** (Z-centric): coordinates that carry the constraints, with $m_{\ell\ell}$, the Z rapidity and the Collins–Soper decay angles as direct target coordinates, exactly invertible back to WRAP.

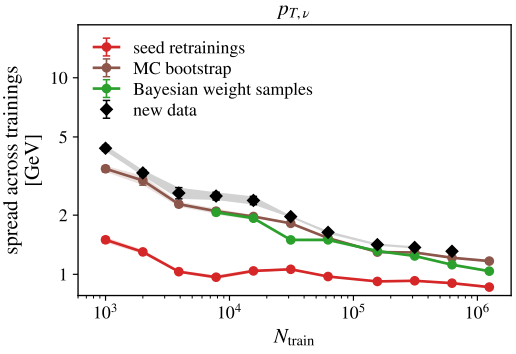
Backup: composition of the band on a spectrum



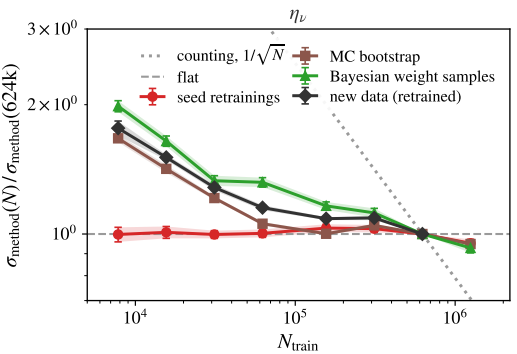
- Percent-level total band
- Data statistics dominate
- Tighter than a counting experiment: each event's posterior spreads over several bins

Backup: the ladder in absolute units

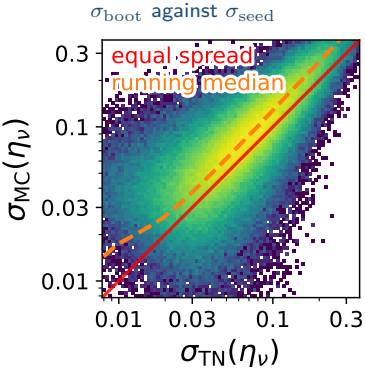
$p_{T,\nu}$

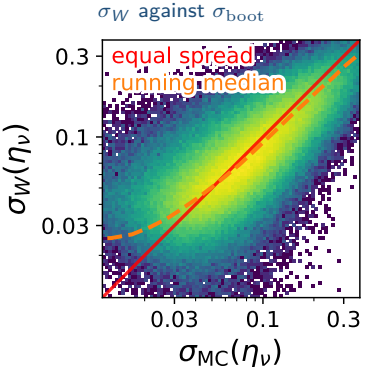


η_ν, normalised to 624k

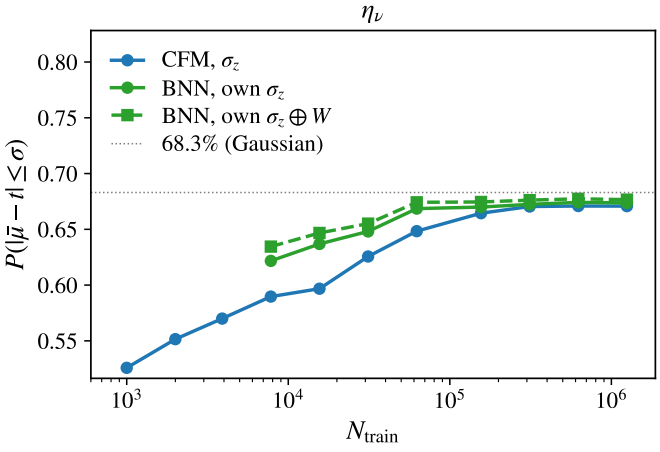


Backup: per-event agreement between the estimators



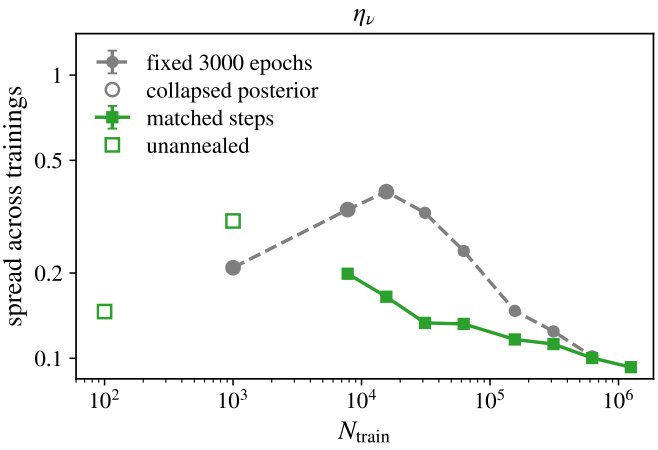


Backup: coverage against truth on the posterior axis



- Posterior width covers the truth
- Training axis: small shift
- Physics sets the scale

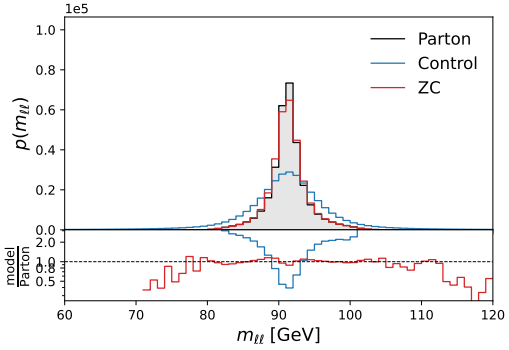
Backup: a training recipe can imitate a data-volume effect



- Same model, two training lengths
- Collapse is the recipe
- Train to the validation optimum, then compare

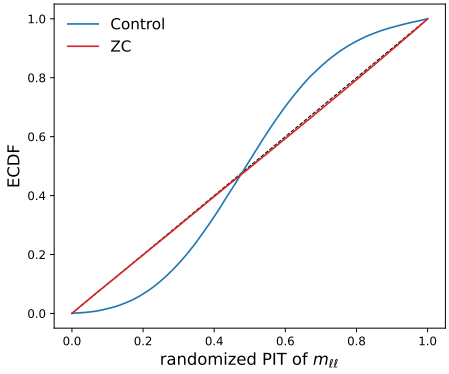
Backup: physics constraints in the coordinates

the Z resonance, unfolded



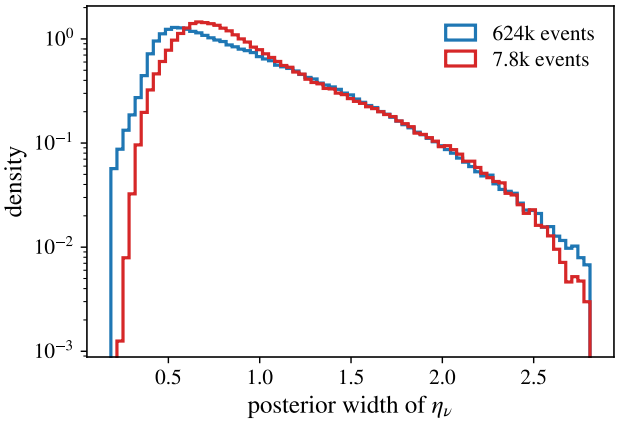
- Standard target: mass implicit
- Z-centric: mass is an axis

is the per-event mass posterior honest?



A straight line means the per-event width is calibrated.
The relations do not move.

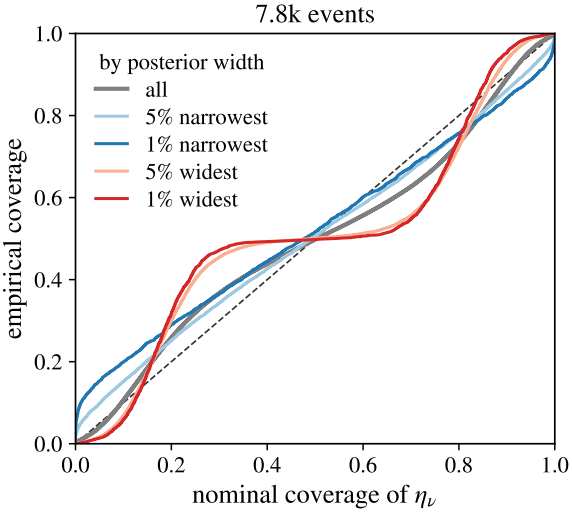
Backup: the posterior-width distribution against training volume



- Width distribution: physics
- Bulk unchanged with data
- Only the already-narrow posteriors sharpen

The hard events stay hard.

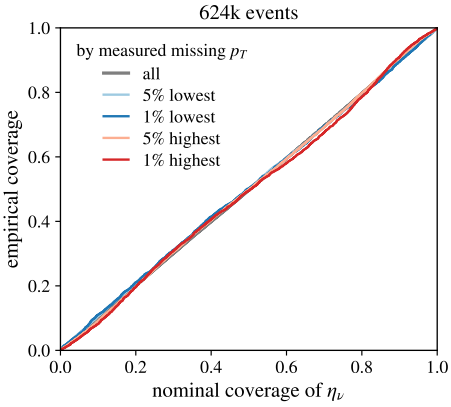
Backup: calibration at low training volume

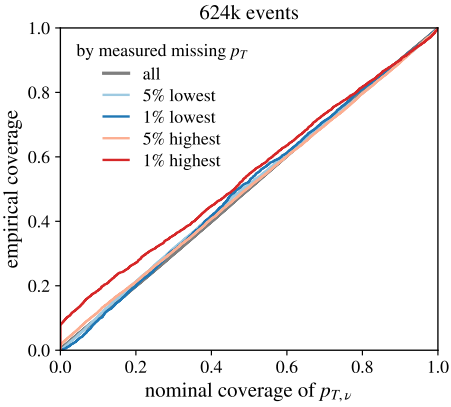


- Two solutions found
- Each lobe too wide
- Narrowest slices over-confident

Local, physical, and absent at the production volume.

Backup: calibration sliced by the measured missing transverse momentum





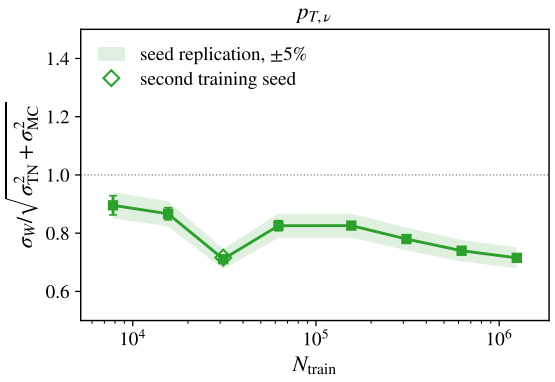
- Same construction as the width slices, selection on a detector-level quantity (legitimate: a function of the record; the truth is not). η_ν calibrated in every slice; $p_{T,\nu}$ calibrated except the 1% with the highest measured missing p_T: 19% of truths in the outer deciles (null 10%), posterior shifted high, the tail where the detector overshoots on the closure slide.

Backup: the new-data comparison, volume by volume

ratio	1k	7.8k	15.6k	31.2k	62.4k	156k	312k	624k
$\sigma_{\mathrm{boot}}/\sigma_{\mathrm{new}}$	0.71–0.97	0.84–0.97	0.83–0.96	0.90–0.99	0.90–0.94	0.89–0.95	0.93–0.99	0.93–1.03
$\sigma_{\mathrm{seed}}/\sigma_{\mathrm{new}}$	0.29–0.41	0.39–0.50	0.44–0.57	0.49–0.65	0.54–0.69	0.61–0.75	0.62–0.75	0.67–0.79

Ranges over the seven non-azimuth observables, pooled rms over events, $B = 8$ up to 15.6k and at 624k, 50 at 31.2k–312k. New-data slices are disjoint from every band and from each other: eight up to 156k, five at 312k, two at 624k. At 7.8k the deficit replicates at a second initialisation (0.90 ± 0.02) and is unchanged when the optimiser stream is re-seeded (1.006 ± 0.005). No deficit on the azimuths with the circular estimator (0.95–1.03).

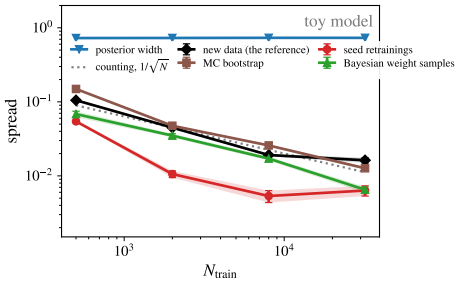
Backup: the Bayesian band and the seed term



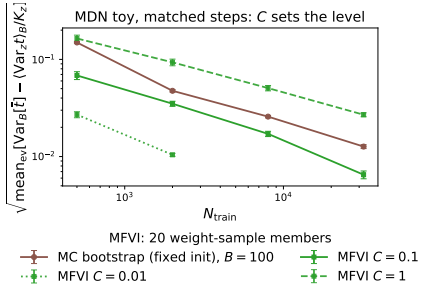
- $\sigma_W/\sigma_{\rm boot} = 0.82$ to 1.24 in rms at every volume, no trend in N (1.07 ± 0.07); three trainings at 31.2k put the training-to-training scatter of σ_W at 3 to 8%.
- $\sigma_W/\sqrt{\sigma_{\rm seed}^2+\sigma_{\rm boot}^2}$ falls from 0.90–1.08 at 7.8k to 0.74–0.82 at 624k, because $\sigma_{\rm seed}/\sigma_{\rm boot}$ rises with N, not because the Bayesian spread moves: a crossing, not a convergence.
- Two Bayesian trainings differ by 1.0 to $1.7\,\sigma_W$ in the posterior mean, so $q(W)$ is not the whole training axis.

Backup: the toy models in detail

small toy, 32 hidden units: the seed floor keeps sinking



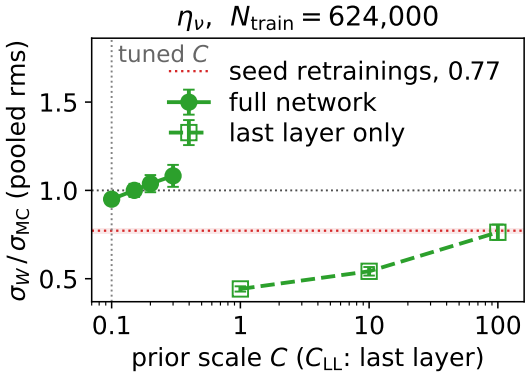
small toy: the prior scale sets the Bayesian level



- Linear-Gaussian regression: the pedestal-corrected estimator recovers injected spread ratios to 1.1%; a mean-field weight posterior gives 0.79 to 0.87 of the exact predictive width.
- Mixture-density toy at matched budget: the fixed-init bootstrap reproduces the new-data spread with median $\sigma_{\text{boot}}/\sigma_{\text{new}} = 1.2$, both falling near $1/\sqrt{N}$; the prior weight C sets the Bayesian level.
- Direction differs from the physics setup (0.83 to 1.03): different model class, matched budget, early stopping. At the validation optimum with the campaign's fixed-initialisation new-data arm the wide toy reads 1.3–1.5 (the fixed-budget figures' new-data arm drew its own initialisations).

Backup: the Bayesian level against the prior scale

η_ν: $\sigma_W/\sigma_{\text{boot}}$ at 624k against the prior scale C



Above the tuned value the prior scale does not set the level: tripling it barely moves the level. A last-layer-only posterior predicts almost the same posterior width, but its weight spread stays well under the bootstrap and keeps climbing with the head's prior scale, so it is an unresolved upper bound rather than a cheap substitute.

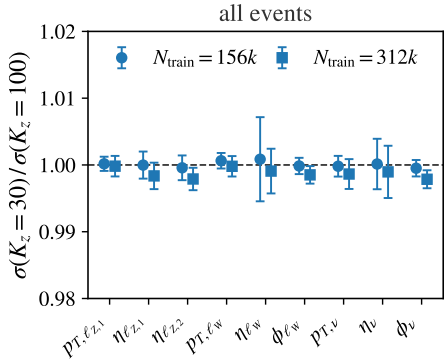
Backup: the controls in numbers (1/2)

control	measured	reading
samples per event: the same 50 weight samples at $K_z = 30$ and 100, 156k and 312k	pooled $\sigma_{30}/\sigma_{100} = 0.998$–$1.001$ (jackknife errors 0.1–0.6%); per width slice 0.994–1.008 (errors $\leq 2.7\%$, max $\lvert z \rvert = 2.9$ over 198 cells); subtracted pedestal 0.3–3.5$\times$ the signal	the subtraction holds to two parts in a thousand pooled and within one percent per slice
optimiser stream re-seeded at fixed init and bootstrap weights (7.8k, $B = 8$)	$\sigma_{\mathrm{boot}}^{\mathrm{reseed}}/\sigma_{\mathrm{boot}} = 1.006 \pm 0.005$ (rms 0.991–1.014)	batch order is not what separates the bootstrap from new data
second initialisation, bootstrap and new data (7.8k, $B = 8$)	$\sigma_{\mathrm{boot}}/\sigma_{\mathrm{new}}$: 0.857–0.966 at the second init against 0.861–0.971 at the first; new-data level $\times 1.031 \pm 0.005$	the deficit replicates observable by observable; a 3% level shift
300 epochs instead of 150, seeds and new data (7.8k, $B = 8$)	seeds $\times$1.30–1.42, new data $\times$1.37–1.56; $\sigma_{\mathrm{seed}}/\sigma_{\mathrm{new}} = 0.39$ against 0.44 (shift 0.94 ± 0.03)	the ratio is schedule-robust, the levels are not; no bootstrap member anneals at 300

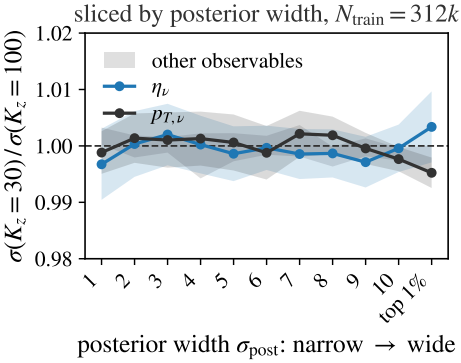
Backup: the controls in numbers (2/2)

control	measured	reading
same seed on an H200, a MIG slice and an A100	weights agree to $\leq 2.4 \times 10^{-7}$; bitwise reproducible on one device	no hardware term and no chaos term in the seed spread
models trained on new against original events (156k new-data band, $3 + 5$ members)	permutation test over all 56 splits: $p = 0.6$–0.8 on $6/7$; η_ν rank $1/56$ at $+4\%$ (all-original control: $+20\%$); a coherent shift $> 0.3\,\sigma_{\mathrm{new}}$ excluded	the generated data is validated at the level of the trained model
two-member new-data band at 624k	relative precision 1.5–4.8% from two-member subsets of the $B = 8$ bands at 62.4k and 156k (disjoint pairs; all subsets 1.6–4.6%); $D_{\mathrm{eff}} \approx 150$–$1800$	usable; the counting error (71%) is the wrong scale
three Bayesian trainings at 31.2k	training-to-training scatter of σ_W: 2.6–8.5% (median 5.7%)	the seed term on the Bayesian level

Backup: the number of posterior samples per event



- Same model, three times the samples



A test that could have failed.

Backup: the finite-sample pedestal and its test

$$\mathrm{Var}_b\big[\bar{t}_b\big] = \underbrace{\mathrm{Var}_W\big[\mathbb{E}_z\, T\big]}_{\sigma^2_{\mathrm{ens}}} + \underbrace{\frac{\mathbb{E}_W\big[\mathrm{Var}_z\, T\big]}{K_z}}_{\text{pedestal}} \qquad \Rightarrow \qquad \hat{\sigma}^2_{\mathrm{ens}} = \mathrm{Var}_b\big[\bar{t}_b\big] - \frac{\widehat{\mathrm{Var}}_z}{K_z}$$

- $\bar{t}_b$: member b's mean over K_z posterior draws of one event; the pedestal is the sampling variance of that mean, subtracted per event, then pooled.
- Test: the same 50 weight samples at $K_z = 30$ and 100 (the 30 draws are the first 30 of the 100; the $K_z = 100$ dump used the gated fixed-step solver, widths agree to 0.1%), 312,286 events, two volumes. Pooled $\hat{\sigma}_{30}/\hat{\sigma}_{100}$: 0.9995–1.0009 (156k), 0.9979–0.9998 (312k), leave-one-member-out errors 0.001–0.006. Per width slice (deciles and the widest 1%, binned on the sibling model's width, no draw shared with the test): 0.994–1.008, errors 0.0006–0.027, max $|z| = 2.9$ over 198 cells.
- Power: the pedestal is 0.5–2.5 times the signal at the median event and 0.3–3.5 times it per slice (largest in the middle azimuth slices); uncorrected, the ratio reads 1.12–1.40 pooled and up to 1.5 per slice. Bound on an error of the pedestal formula: about 1% pooled, a few % per slice. An event bootstrap understates the errors 2–13× (sampling noise is coherent across events within a weight sample).
- K_z-invariant: pooled rms and slice means. Not invariant: per-event medians, kept fractions, rank agreements; compare those across K_z only after adding noise of variance $\widehat{\mathrm{Var}}_z\,(1/K_{\mathrm{low}} - 1/K_{\mathrm{high}})$ to the member means (verified to 0.3%).
- A per-event clip at zero before pooling biases the low-K_z side up (η_ν at 312k: 1.012 clipped, 0.999 unclipped).

Backup: the 1.25M rung, doubling the data above the production volume

- 1,248,000 training events: the original split plus newly generated events, same recipe (150 epochs, so the optimiser steps double with the data); seed and bootstrap bands of $B = 8$ at $K_z = 30$.
- Spread ratio $\sigma(1.25\mathrm{M})/\sigma(624\mathrm{k})$, pooled rms, seven non-azimuth observables: seeds 0.94–1.03, bootstrap 0.95–1.03 (member errors 2–4% and 1–2%), Bayesian weight samples 0.88–0.92 (50 draws, one training). The new-data reference's own step from 312k to 624k was 0.89–1.00.
- Accuracy: the paired per-member test error improves by 0.5–2% (0.2–0.4% in rms); η_ν degrades by 0.9–1.3% ($t = 22$–44), the one watch item, and the one observable with a weak group signal in the permutation test.
- Reading: a plateau. Above about 600k events neither the accuracy nor the spread of this recipe responds to more data. Not an attribution: data and steps doubled together, and the one fixed-volume budget intervention (300 epochs at 7.8k) moves the spread the other way.
- Azimuths are unmeasured at $K_z = 30$ here (the pedestal exceeds the signal). The Bayesian band contracts faster than the two retraining bands over this doubling, as over the whole ladder; a second Bayesian training is in flight.

Backup: related work

- **OmniFold**: Andreassen, Komiske, Metodiev, Nachman, Thaler, Phys. Rev. Lett. **124** (2020) 182001, arXiv:1911.09107. Iterative reweighting, unbinned and simultaneous in all observables.
- **Generative unfolding with invertible networks**: Bellagente et al., SciPost Phys. **9** (2020) 074, arXiv:2006.06685; per-event posteriors from a conditional INN.
- **Uncertainties of generative networks**: Bellagente, Haussmann, Luchmann, Plehn, SciPost Phys. **13** (2022) 003, arXiv:2104.04543.
- **Butter, Diefenbacher, Plehn, Vogel**, arXiv:2608.21509: learned systematic and statistical uncertainties for *generation*, calibrated against an explicit toy likelihood. Complementary: their axes are learned by the loss on a density, ours are measured by repeated training on an inverse problem, against new-data retraining.