Robust Bayesian tuning of hadronisation models

outliers, surrogates and gradient-based MCMC

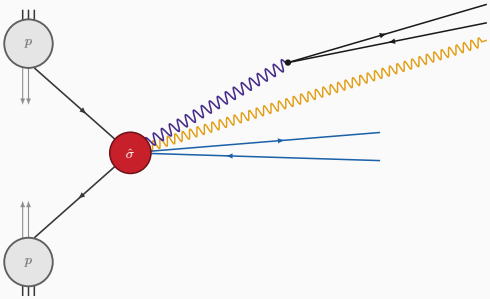
James Whitehead
Jagiellonian University, Kraków

Bohr seminar series · University of Manchester
13 November 2026



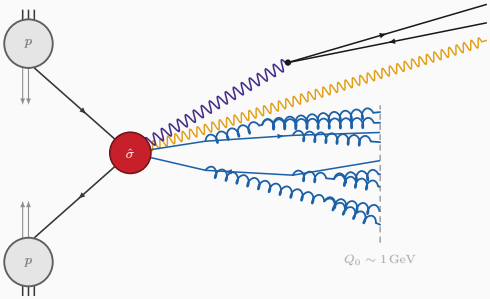
JAGIELLONIAN UNIVERSITY
IN KRAKÓW

Anatomy of a simulated event



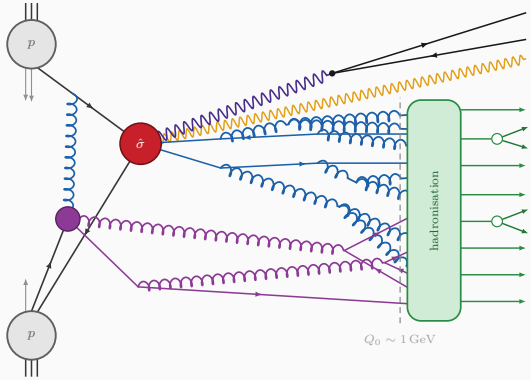
1. **hard process**: up to (N)NLO in α_s

Anatomy of a simulated event



1. **hard process**: up to (N)NLO in α_s
2. **parton shower**: approximate evolution, down to $Q_0 \sim 1\,\mathrm{GeV}$

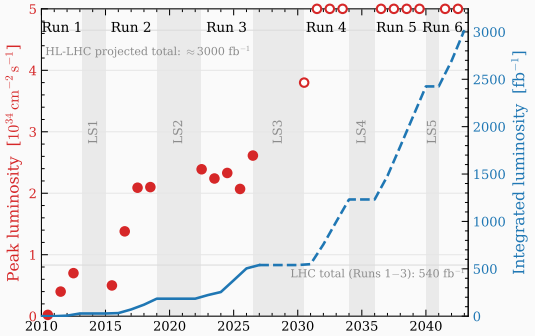
Anatomy of a simulated event



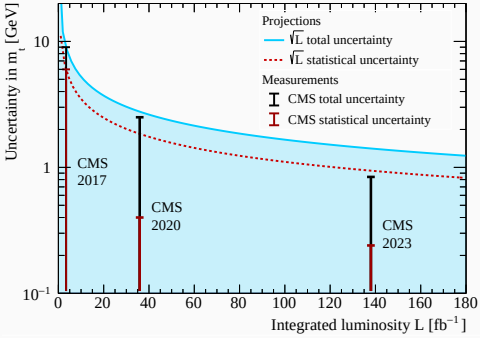
1. **hard process**: up to (N)NLO in α_s
2. **parton shower**: approximate evolution, down to $Q_0 \sim 1\,\mathrm{GeV}$
3. **hadronisation**, MPI, decays: semi-empirical models

Below $\sim 1\,\mathrm{GeV}$, (currently) no first-principles calculation: rely on models with fitted parameters

More data, more precision?



[LHC luminosity, recorded and projected]



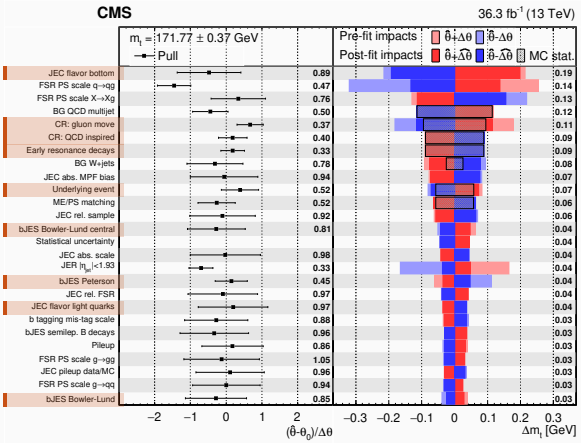
[m_t at CMS: uncertainty falls *faster* than $1/\sqrt{L}$]

Experimental uncertainties shrink with L; pure theory uncertainties don't.

Modelling uncertainties can shrink, if new constraints lead to model improvement.

[Belvedere et al. (2402.07985)]

Why tuning matters: the error budget



[CMS m_t ranked nuisance impacts; shaded: non-perturbative modelling]

Top-quark mass (CMS 2023): largest impact is roughly *string vs. cluster fragmentation*; CR models next.

α_s **from** e^+e^- (PDG 2024): *"dominant uncertainty is the hadronization uncertainty"*.

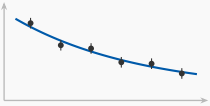
In every analysis: models enter parton-to-hadron *unfolding*.

After the perturbative 'NNLO revolution', non-perturbative modelling is increasingly a leading uncertainty

[CMS (2302.01967), Fig. 7, Table 2, PDG 2024, QCD review, §9.4.4, ATLAS, ATL-PHYS-PUB-2014-021 (A14)]

What do we want from model calibration?

1 A meaningful central prediction

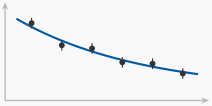


Well-motivated model parameters that can describe a range of data, preferably moderately insensitive to tune methodology or data choices, weights.

→ nominal MC for analyses

What do we want from model calibration?

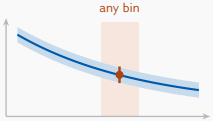
1 A meaningful central prediction



Well-motivated model parameters that can describe a range of data, preferably moderately insensitive to tune methodology or data choices, weights.

→ nominal MC for analyses

2 Uncertainties on predictions

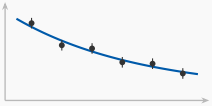


Propagation of model/parameter uncertainties through to future observables.

→ the modelling systematic of every measurement

What do we want from model calibration?

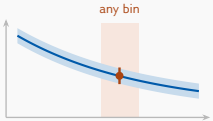
1 A meaningful central prediction



Well-motivated model parameters that can describe a range of data, preferably moderately insensitive to tune methodology or data choices, weights.

→ nominal MC for analyses

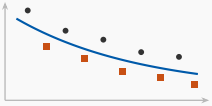
2 Uncertainties on predictions



Propagation of model/parameter uncertainties through to future observables.

→ the modelling systematic of every measurement

3 A guide to model inadequacy



Which data the model cannot describe, and which data sets it cannot describe simultaneously.

→ where to improve the model; where not to trust the tune

Without rigorous uncertainties, we cannot interpret comparisons with data.

Same problems, different names

	PDF fits	MC tuning	flavour, EW, PDG	heavy ions, cosmology	statistics
data and model disagree	tolerance $\Delta\chi^2 \gg 1$	hand-set weights, dropped bins	Rfit ranges; scale factor S	one model σ, or none	model discrepancy; robust M-estimators

[Huber (1964), Sacks et al. (1989), Kennedy and O'Hagan (2001), Hocker et al. (Rfit), Navas et al. (PDG 2024, §5.2.2)]

Same problems, different names

	PDF fits	MC tuning	flavour, EW, PDG	heavy ions, cosmology	statistics
data and model disagree	tolerance $\Delta\chi^2 \gg 1$	hand-set weights, dropped bins	Rfit ranges; scale factor S	one model σ, or none	model discrepancy; robust M-estimators
model too slow	interpolation grids	polynomials, rationals		GP, NN emulators	GP surrogates of computer codes

[Huber (1964), Sacks et al. (1989), Kennedy and O'Hagan (2001), Hocker et al. (Rfit), Navas et al. (PDG 2024, §5.2.2)]

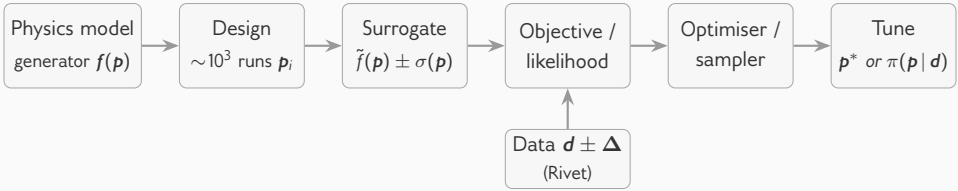
Same problems, different names

	PDF fits	MC tuning	flavour, EW, PDG	heavy ions, cosmology	statistics
data and model disagree	tolerance $\Delta\chi^2 \gg 1$	hand-set weights, dropped bins	Rfit ranges; scale factor S	one model σ, or none	model discrepancy; robust M-estimators
model too slow	interpolation grids	polynomials, rationals		GP, NN emulators	GP surrogates of computer codes
point estimate or distribution?	Hessian vs. replicas	eigentunes, 'Var' tunes	frequentist vs. Bayesian	MCMC	posterior sampling

Model calibration is a general problem disguised as a specific one; diferent approaches per sub-field.

[Huber (1964), Sacks et al. (1989), Kennedy and O'Hagan (2001), Hocker et al. (Rfit), Navas et al. (PDG 2024, §5.2.2)]

What does it mean to 'tune' a model?

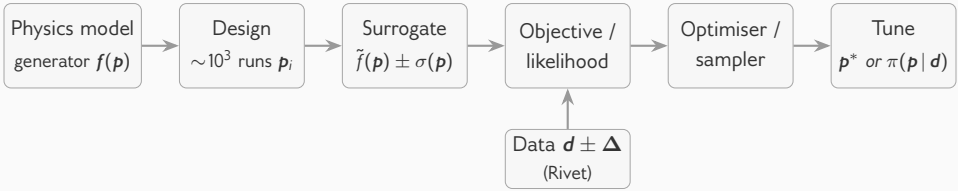


- $d \sim 10$–100 generator parameters $\boldsymbol{p}$ without first-principles values.
- **Tuning**: choose $\boldsymbol{p}$ so that $\boldsymbol{f}(\boldsymbol{p})$ matches $N \sim 10^3$–10^4 data values (histogram bins) $\boldsymbol{d} \pm \boldsymbol{\Delta}$.
- The MC generator is slow: use a fast *surrogate*.

This talk: one answer to each problem

[T. Procter, A. Siódmok, JW, *Robust Bayesian calibration of hadronisation model parameters using MCMC methods*, in preparation]

What does it mean to 'tune' a model?



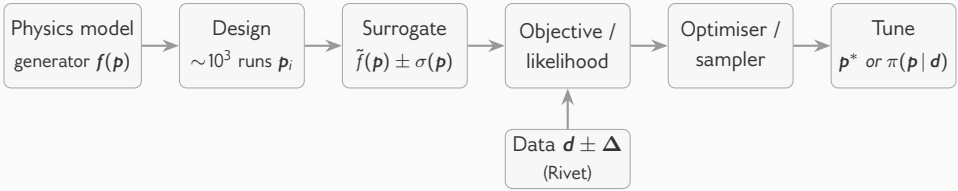
- $d \sim 10$–100 generator parameters $\boldsymbol{p}$ without first-principles values.
- **Tuning**: choose $\boldsymbol{p}$ so that $\boldsymbol{f}(\boldsymbol{p})$ matches $N \sim 10^3$–10^4 data values (histogram bins) $\boldsymbol{d} \pm \boldsymbol{\Delta}$.
- The MC generator is slow: use a fast *surrogate*.

This talk: one answer to each problem

1. a robust likelihood: outliers expected, not fatal;

[T. Procter, A. Siódmok, JW, *Robust Bayesian calibration of hadronisation model parameters using MCMC methods*, in preparation]

What does it mean to 'tune' a model?



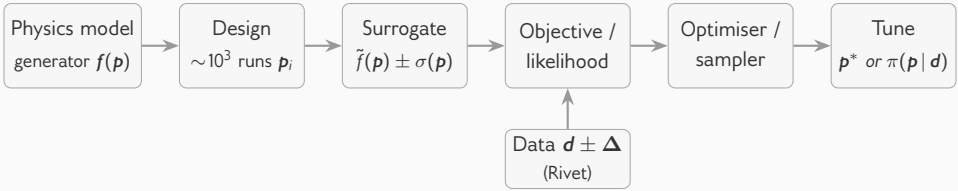
- $d \sim 10$–100 generator parameters $\boldsymbol{p}$ without first-principles values.
- **Tuning**: choose $\boldsymbol{p}$ so that $\boldsymbol{f}(\boldsymbol{p})$ matches $N \sim 10^3$–10^4 data values (histogram bins) $\boldsymbol{d} \pm \boldsymbol{\Delta}$.
- The MC generator is slow: use a fast *surrogate*.

This talk: one answer to each problem

1. a robust likelihood: outliers expected, not fatal;
2. a calibrated surrogate that knows its own uncertainty;

[T. Procter, A. Siódmok, JW, *Robust Bayesian calibration of hadronisation model parameters using MCMC methods*, in preparation]

What does it mean to 'tune' a model?



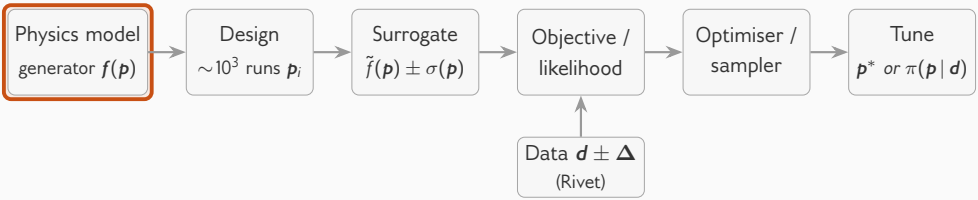
- $d \sim 10$–100 generator parameters $\boldsymbol{p}$ without first-principles values.
- **Tuning**: choose $\boldsymbol{p}$ so that $\boldsymbol{f}(\boldsymbol{p})$ matches $N \sim 10^3$–10^4 data values (histogram bins) $\boldsymbol{d} \pm \boldsymbol{\Delta}$.
- The MC generator is slow: use a fast *surrogate*.

This talk: one answer to each problem

1. a robust likelihood: outliers expected, not fatal;
2. a calibrated surrogate that knows its own uncertainty;
3. gradient-based MCMC: obtain a posterior distribution, not a point estimate.

[T. Procter, A. Siódmok, JW, *Robust Bayesian calibration of hadronisation model parameters using MCMC methods*, in preparation]

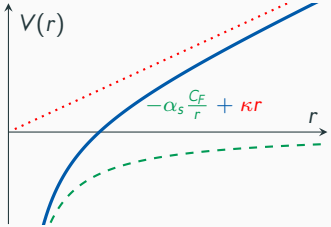
I. Hadronisation and soft QCD



From partons to hadrons: confinement

Lattice QCD: the static $q\bar{q}$ potential grows linearly, $V(r) \sim \kappa\, r$, with $\kappa \approx 1\,\mathrm{GeV/fm}$.

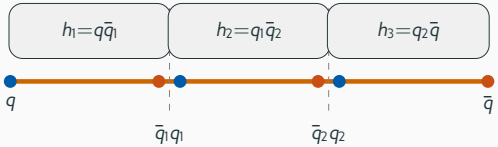
- Short distance: asymptotic freedom $\Rightarrow$ pQCD.
- Long distance: *colour flux tube*, string of tension κ.
- String breaks when creating $q\bar{q}$ pair is cheaper.



Coulomb + linear 'Cornell' potential.

QCD tells us *what* happens qualitatively, but not the details event-by-event $\longrightarrow$ models.

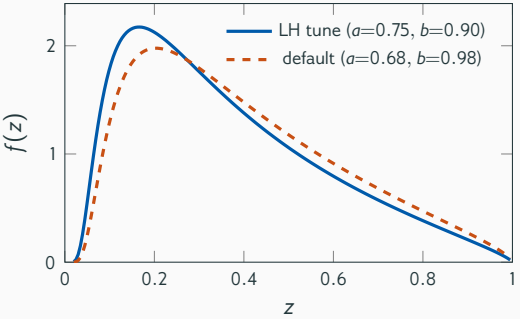
The Lund string model



The string breaks by $q\bar{q}$ pair creation (tunnelling):

$$\mathcal{P} \propto \exp\Big(-\frac{\pi m_\perp^2}{\kappa}\Big), \quad m_\perp^2 = m^2 + p_\perp^2$$

- Gaussian $p_\perp$, width σ;
- flavour hierarchy: s and diquarks suppressed.



Each hadron takes a fraction z of the remaining light-cone momentum:

$$f(z) \propto \frac{1}{z}(1-z)^a \exp\Big(-\frac{b\, m_\perp^2}{z}\Big)$$

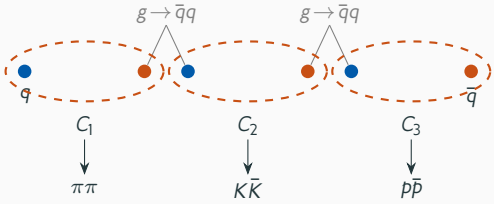
Parameters: a, b, σ, heavy-quark $r_{c,b}$, and ∼10 flavour probabilities.

Default in Pythia.

[Andersson et al. (1983), Sjostrand (1986), Bowler (1981)]

The cluster model

1. At the shower cutoff, split every gluon, $g \to q\bar{q}$.
2. Colour-connect neighbours into clusters: low masses, independent of the hard scale (*preconfinement*).
3. Clusters decay to two hadrons; heavy ones first *fission*.



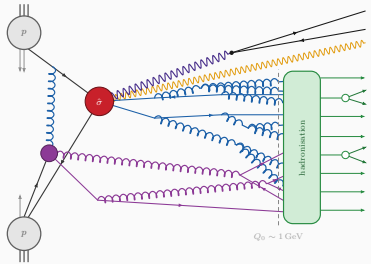
Fewer parameters than string model, but cruder kinematics.

Default in Herwig *and* Sherpa.

Two families: both physically-motivated & consistent with data, for some model parameters **p**.

[Webber (1984), Amati and Veneziano (1979), Gieseke et al. (1710.10906), Hoang et al. (2404.09856)]

More than one collision: MPI, underlying event, colour reconnection



secondary scatterings between the proton remnants

Multiple parton interactions: several scatterings per collision ⇒ underlying event (UE).

- regulate $p_\perp \to 0$: $p_\perp^{\min}(s) = p_{\perp,0}^{\min}\left(\frac{b+\sqrt{s}}{E_0}\right)^c$;
- matter overlap μ^2; soft ladders N_{ladder}, B_{ladder}; diffraction R_{diff}.

Colour reconnection: reassign colour labels to shorten colour strings (m_0, m_{pseudo}, `junctionCorrection`).

Measured transverse to the leading track, at 0.9–13 TeV.

10 parameters set the distribution of soft activity in every collision

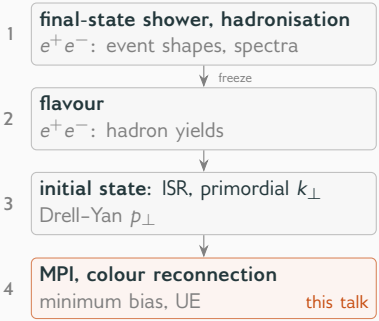
[Bellm et al. (1911.13149), Christiansen and Skands (1505.01681), Bierlich et al. (2203.11601)]

Event Generator tuning in practice

Standard workflow (Professor)

1. run the generator at $\mathcal{O}(10^3)$ points in a box;
2. fit a polynomial per bin: the surrogate;
3. minimise a weighted χ^2;
4. uncertainty: *eigentunes* (diagonalise Hessian)

A typical staged tune



[Buckley et al. (0907.2973), Krishnamoorthy et al. (2103.05748), Divisova et al. (2509.02348)]

Event Generator tuning in practice

Standard workflow (Professor)

1. run the generator at $\mathcal{O}(10^3)$ points in a box;
2. fit a polynomial per bin: the surrogate;
3. minimise a weighted χ^2;
4. uncertainty: *eigentunes* (diagonalise Hessian)

Why in stages?

- *Curse of dimensionality*: many parameters against many distributions at once is beyond these methods.
- Split by *heuristics* (which data probe which component), not routinely tested.
- In principle a cycle: re-tune stage 1 with stage 4 frozen, etc.

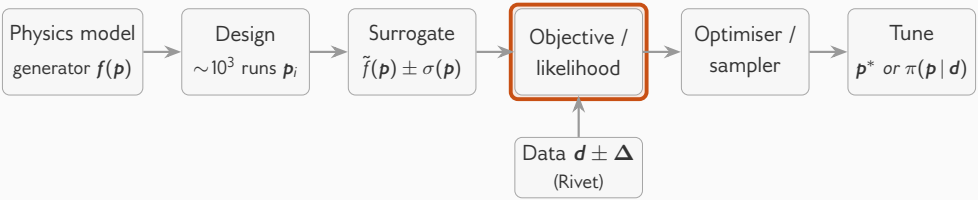
Each stage absorbs everything frozen before it.

A typical staged tune



[Buckley et al. (0907.2973), Krishnamoorthy et al. (2103.05748), Divisova et al. (2509.02348)]

II. Outliers and robust statistics



Tuning as minimisation

Standardised residual of bin b:

$$r_b(\boldsymbol{p}) = \frac{f_b(\boldsymbol{p}) - d_b}{\Delta_b}$$

Goodness of fit and tune:

$$\chi^2(\boldsymbol{p}) = \sum_b w_b^2 \, r_b^2(\boldsymbol{p}), \qquad \boldsymbol{p}^* = \operatorname*{argmin}_{\boldsymbol{p}} \chi^2(\boldsymbol{p})$$

Tuning as minimisation

Standardised residual of bin b:

$$r_b(\boldsymbol{p}) = \frac{f_b(\boldsymbol{p}) - d_b}{\Delta_b}$$

Goodness of fit and tune:

$$\chi^2(\boldsymbol{p}) = \sum_b w_b^2\, r_b^2(\boldsymbol{p}), \qquad \boldsymbol{p}^* = \operatorname*{argmin}_{\boldsymbol{p}} \chi^2(\boldsymbol{p})$$

Statistical reading: if $d_b \sim \mathcal{N}\big(f_b(\boldsymbol{p}), \Delta_b^2\big)$ independently, then

$$\chi^2(\boldsymbol{p}) = -2 \log \pi(\boldsymbol{d} \mid \boldsymbol{p}) + \text{const.}$$

Weights act as *replication*: bin b counted w_b^2 times.

Hidden assumption

A 'true' $\boldsymbol{p}$ exists:
data $\approx$ model + Gaussian noise.
Then $\chi^2/N \approx 1$ at $\mathbf{p}^*$.

Tuning as minimisation

Standardised residual of bin b:

$$r_b(\boldsymbol{p}) = \frac{f_b(\boldsymbol{p}) - d_b}{\Delta_b}$$

Goodness of fit and tune:

$$\chi^2(\boldsymbol{p}) = \sum_b w_b^2\, r_b^2(\boldsymbol{p}), \qquad \boldsymbol{p}^* = \operatorname*{argmin}_{\boldsymbol{p}} \chi^2(\boldsymbol{p})$$

Statistical reading: if $d_b \sim \mathcal{N}\big(f_b(\boldsymbol{p}), \Delta_b^2\big)$ independently, then

$$\chi^2(\boldsymbol{p}) = -2 \log \pi(\boldsymbol{d} \mid \boldsymbol{p}) + \text{const.}$$

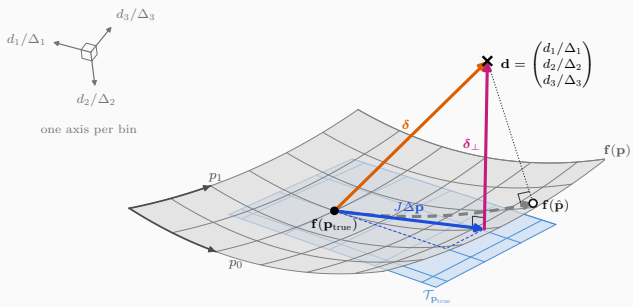
Weights act as *replication*: bin b counted w_b^2 times.

χ^2 minimisation alone gives at most one of the three wants.

Against the three wants (slide 5)

1. central prediction: yes, but $\chi^2/N \gg 1$ from a few bins: meaningful?
2. meaningful uncertainties: no; $\Delta\chi^2 = 1$ relies on the assumption;
3. model inadequacy: not really; χ^2/N says 'inconsistent overall'

Large residuals are not (actually) a surprise



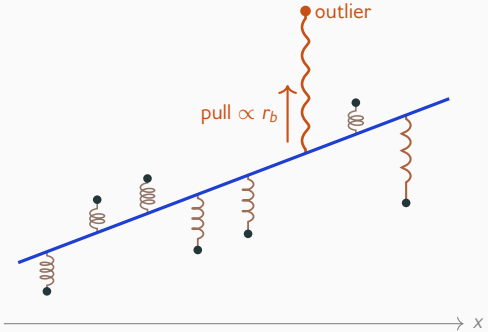
[$N = 3$ bins, 2 parameters; units of Δ_b]

Large residuals are *expected*: the traces of model inadequacy

- $d \approx 10$ parameters span a d-dimensional *model manifold* among $N \approx 3600$ bins.
- Mismodelling $\boldsymbol{\delta}$ splits into a tangential part, absorbed by $\boldsymbol{p}$, and an orthogonal part ($N - d$ dimensions): irreducible.

$\Rightarrow$ Some bins cannot be fitted for *any* $\boldsymbol{p}$

χ^2: large residuals have unbounded pull



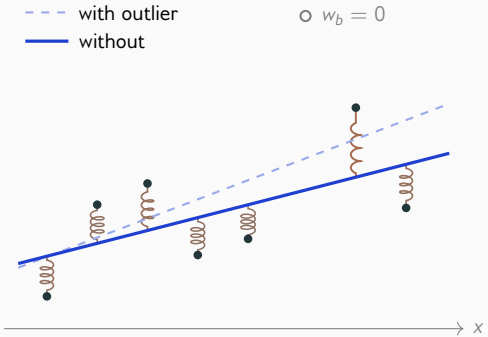
χ^2 as springs: stiffness $\propto w_b^2/\Delta_b^2$ (Hooke's law)

$$\frac{\partial \chi^2}{\partial p_i} = 2 \sum_b \frac{w_b^2}{\Delta_b} \underbrace{r_b}_{\text{pull}} \frac{\partial f_b}{\partial p_i}$$

- Pull grows *linearly* with the residual: one precise, unfittable bin drags all parameters.
- Usual fix: set $w_b = 0$ by hand, drop bins outside surrogate's; binary decisions with an arbitrary threshold.

[Krishnamoorthy et al. (2103.05748)]

χ^2: large residuals have unbounded pull



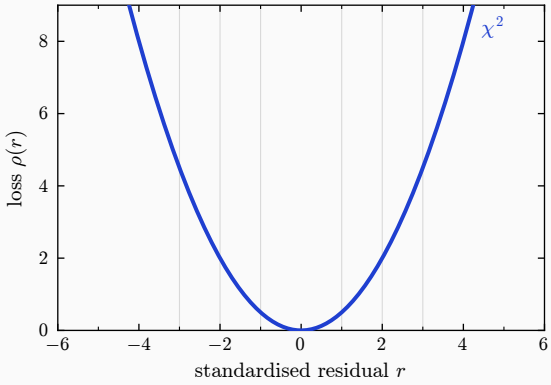
χ^2 as springs: stiffness $\propto w_b^2/\Delta_b^2$ (Hooke's law)

$$\frac{\partial \chi^2}{\partial p_i} = 2 \sum_b \frac{w_b^2}{\Delta_b} \underbrace{r_b}_{\text{pull}} \frac{\partial f_b}{\partial p_i}$$

- Pull grows *linearly* with the residual: one precise, unfittable bin drags all parameters.
- Usual fix: set $w_b = 0$ by hand, drop bins outside surrogate's; binary decisions with an arbitrary threshold.

[Krishnamoorthy et al. (2103.05748)]

Robust losses

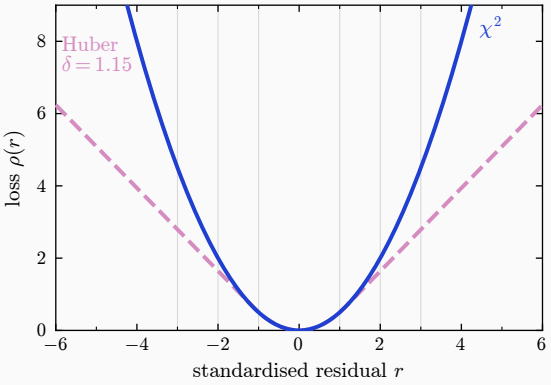


Generalise χ^2 to a per-bin loss ρ:

$$\mathcal{L}_\rho(\boldsymbol{p}) = 2\sum_b w_b^2\,\rho\big(r_b(\boldsymbol{p})\big), \qquad \rho_{\chi^2} = \tfrac{1}{2}r^2$$

[Huber (1964), Maronna et al. (2019), Lange et al. (1989)]

Robust losses



Generalise χ^2 to a per-bin loss ρ:

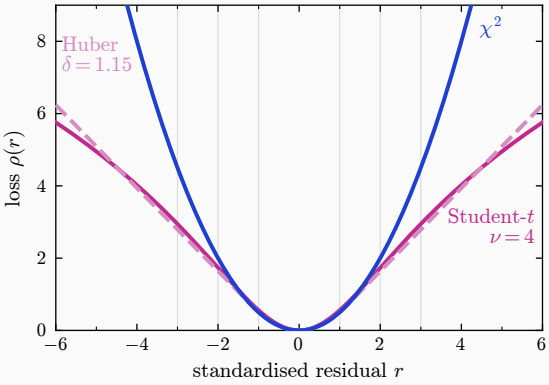
$$\mathcal{L}_\rho(\boldsymbol{p}) = 2\sum_b w_b^2\,\rho\big(r_b(\boldsymbol{p})\big), \qquad \rho_{\chi^2} = \tfrac{1}{2}r^2$$

- **Huber**: quadratic core, *linear* beyond δ

$$\rho_{\mathrm{H}} = \begin{cases} \tfrac{1}{2}r^2 & |r| \le \delta \\ \delta|r| - \tfrac{1}{2}\delta^2 & |r| > \delta \end{cases}$$

[Huber (1964), Maronna et al. (2019), Lange et al. (1989)]

Robust losses



Generalise χ^2 to a per-bin loss ρ:

$$\mathcal{L}_\rho(\boldsymbol{p}) = 2\sum_b w_b^2\,\rho\big(r_b(\boldsymbol{p})\big), \qquad \rho_{\chi^2} = \tfrac{1}{2}r^2$$

- **Huber**: quadratic core, *linear* beyond δ

$$\rho_{\mathrm{H}} = \begin{cases} \tfrac{1}{2}r^2 & |r| \le \delta \\ \delta|r| - \tfrac{1}{2}\delta^2 & |r| > \delta \end{cases}$$

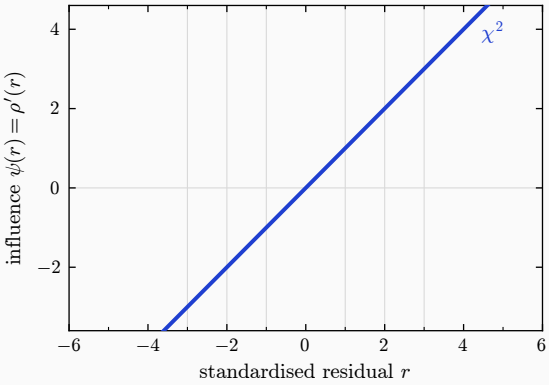
- **Student-t**: *logarithmic* growth

$$\rho_\nu = \frac{\nu+1}{2}\log\Big(1+\frac{r^2}{\nu}\Big)$$

Robust regression (Huber, 1964): estimators that degrade gracefully when a fraction ε of the data is 'contaminated'.

[Huber (1964), Maronna et al. (2019), Lange et al. (1989)]

Influence functions: how hard can one bin pull?



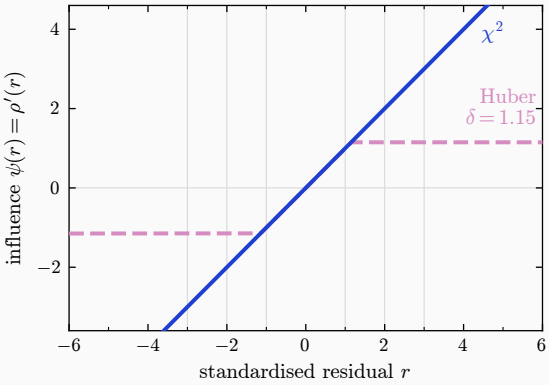
The pull of one bin is the *influence function* $\psi = \rho'$:

$$\frac{\partial \mathcal{L}_\rho}{\partial p_i} = 2 \sum_b \frac{w_b^2}{\Delta_b} \psi(r_b) \frac{\partial f_b}{\partial p_i}$$

- χ^2: *unbounded* (Hookean spring);

[Hampel (1974)]

Influence functions: how hard can one bin pull?



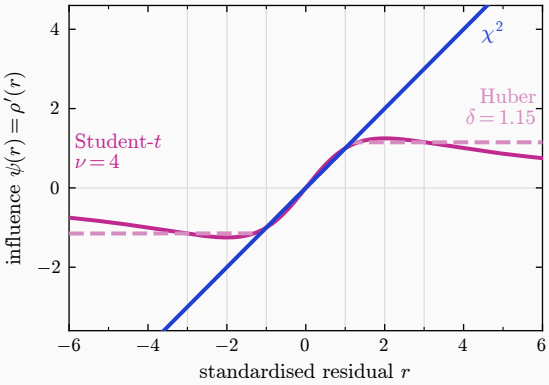
The pull of one bin is the *influence function* $\psi = \rho'$:

$$\frac{\partial \mathcal{L}_\rho}{\partial p_i} = 2 \sum_b \frac{w_b^2}{\Delta_b} \psi(r_b) \frac{\partial f_b}{\partial p_i}$$

- χ^2: *unbounded* (Hookean spring);
- Huber: *bounded* (spring yields at δ);

[Hampel (1974)]

Influence functions: how hard can one bin pull?



The pull of one bin is the *influence function* $\psi = \rho'$:

$$\frac{\partial \mathcal{L}_\rho}{\partial p_i} = 2 \sum_b \frac{w_b^2}{\Delta_b} \psi(r_b) \frac{\partial f_b}{\partial p_i}$$

- χ^2: *unbounded* (Hookean spring);
- Huber: *bounded* (spring yields at δ);
- Student-t: *redescending* (deforms/slackens)

Redescending: smooth, automatic $w_b \to 0$.
Cost: non-convex, so several minima are possible.

[Hampel (1974)]

Huber's optimality: minimax theorem

Theorem (Huber 1964)

If the residuals are Gaussian except for a fraction ε of *arbitrary* (symmetric) contamination, the Huber loss has the smallest *worst-case* variance of all M-estimators, for $\delta(\varepsilon)$ given by:

$$\frac{2\phi(\delta)}{\delta} - 2\Phi(-\delta) = \frac{\varepsilon}{1-\varepsilon}.$$

Matched (δ, ε) pairs

δ	2.0	1.345	1.15	1.0
ε	0.8%	5.8%	9.7%	14%

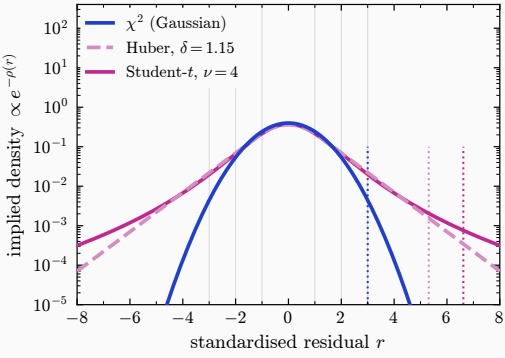
Idea: worst case is Gaussian core, exponential tails

For $\delta = 1.15$

- bounded influence; worst-case variance 1.48 (χ^2: ∞)
- 'insurance premium': 7% efficiency lost on Gaussian data
- convex $\Rightarrow$ a unique minimum

[Huber (1964), Huber and Ronchetti (2009, ch. 4), Maronna et al. (2019)]

Robust loss functions → robust likelihoods



implied density $\pi(r) \propto e^{-\rho(r)}$;
dotted: tail as rare as Gaussian 3σ

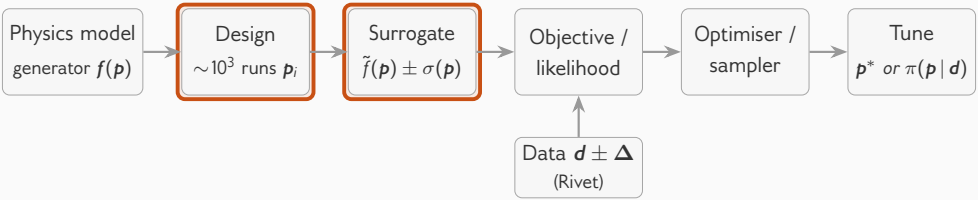
Fat tails: large residuals are *expected*, not vanishingly rare (hence decisive).

likelihood	$\lvert r\rvert$ as rare as Gaussian 3σ
Gaussian	3.0
Huber, $\delta = 1.15$	5.3
Student-t, $\nu = 6\,/\,4\,/\,3$	4.9 / 6.6 / 9.2

Matched by efficiency on 'clean' Gaussian data (ARE):

ARE	0.95	0.925	0.90
Huber δ	1.345	1.15	1.0
Student-t ν	6	4	3

III. Surrogates



Surrogates for expensive calculations

The cost problem

- One parameter point $\mathbf{p}$ = full generation and analysis (at four energies)
- Must evaluate as many points as are required by the minimisation/sampling algorithm before convergence ($\gtrsim 10^6$ for MCMC)

$\Rightarrow$ Run $\sim 10^3$ points; fit a cheap per-bin surrogate $\tilde{f}_b(\boldsymbol{p})$.

On the market: polynomials (Professor), Padé rationals (Apprentice), GPs, NNs.

Our requirements

[Buckley et al. (0907.2973), Krishnamoorthy et al. (2103.05748), Ilten et al. (1610.08328), Lazzarin et al. (2010.02213), Bradbury et al. (2018)]

Surrogates for expensive calculations

The cost problem

- One parameter point $\mathbf{p}$ = full generation and analysis (at four energies)
- Must evaluate as many points as are required by the minimisation/sampling algorithm before convergence ($\gtrsim 10^6$ for MCMC)

$\Rightarrow$ Run $\sim 10^3$ points; fit a cheap per-bin surrogate $\tilde{f}_b(\boldsymbol{p})$.

On the market: polynomials (Professor), Padé rationals (Apprentice), GPs, NNs.

Our requirements

1. **locally refinable**: no assumptions about global shape

[Buckley et al. (0907.2973), Krishnamoorthy et al. (2103.05748), Ilten et al. (1610.08328), Lazzarin et al. (2010.02213), Bradbury et al. (2018)]

Surrogates for expensive calculations

The cost problem

- One parameter point $\mathbf{p}$ = full generation and analysis (at four energies)
- Must evaluate as many points as are required by the minimisation/sampling algorithm before convergence ($\gtrsim 10^6$ for MCMC)

$\Rightarrow$ Run $\sim 10^3$ points; fit a cheap per-bin surrogate $\tilde{f}_b(\boldsymbol{p})$.

On the market: polynomials (Professor), Padé rationals (Apprentice), GPs, NNs.

Our requirements

1. **locally refinable**: no assumptions about global shape
2. a **calibrated uncertainty** $\sigma_b(\boldsymbol{p})$

[Buckley et al. (0907.2973), Krishnamoorthy et al. (2103.05748), Ilten et al. (1610.08328), Lazzarin et al. (2010.02213), Bradbury et al. (2018)]

Surrogates for expensive calculations

The cost problem

- One parameter point $\mathbf{p}$ = full generation and analysis (at four energies)
- Must evaluate as many points as are required by the minimisation/sampling algorithm before convergence ($\gtrsim 10^6$ for MCMC)

$\Rightarrow$ Run $\sim 10^3$ points; fit a cheap per-bin surrogate $\tilde{f}_b(\boldsymbol{p})$.

On the market: polynomials (Professor), Padé rationals (Apprentice), GPs, NNs.

Our requirements

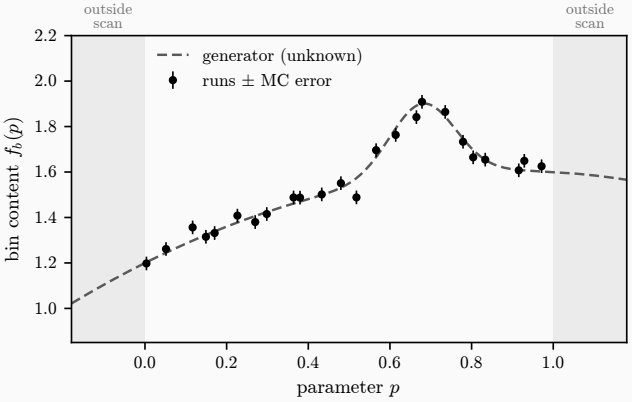
1. **locally refinable**: no assumptions about global shape
2. a **calibrated uncertainty** $\sigma_b(\boldsymbol{p})$
3. **differentiable** (JAX): cheap gradients

A global polynomial meets none of these.

With a calibrated surrogate uncertainty estimate, the surrogate itself can enter into the statistical model/likelihood.

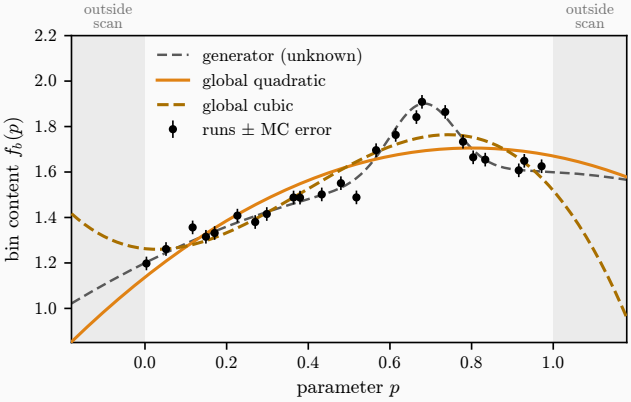
[Buckley et al. (0907.2973), Krishnamoorthy et al. (2103.05748), Ilten et al. (1610.08328), Lazzarin et al. (2010.02213), Bradbury et al. (2018)]

Toy illustration (1d)



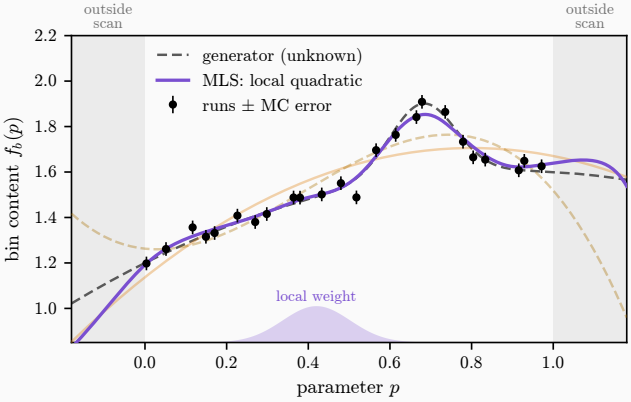
Generator runs, with Monte Carlo noise.
True response $f_b(p)$ is unknown.

Toy illustration (1d)



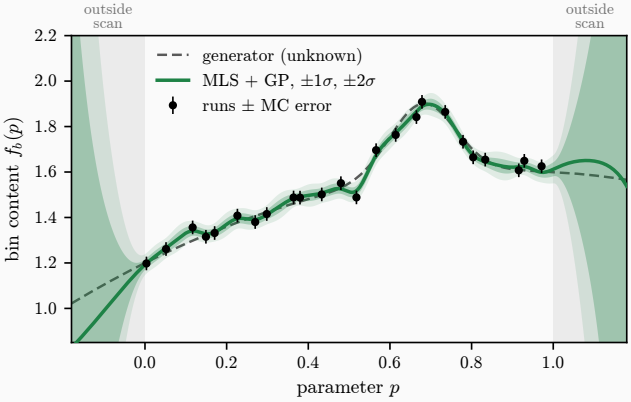
Global polynomial (Professor-style): biased near features, weighted by sampling density, erratic outside the scan, no error estimate.

Toy illustration (1d)



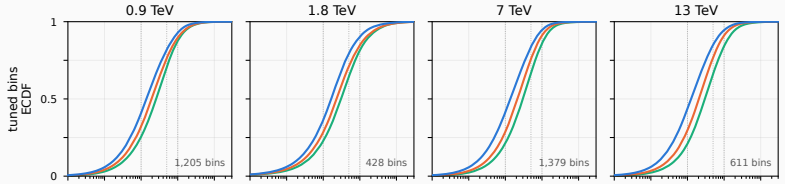
Moving least squares: a local quadratic refitted at every p, downweighting distant runs/parameter points. Refinable by denser subsampling.

Toy illustration (1d)



MLS + GP on leave-one-out residuals: corrects the mean and gives $\sigma(p)$ (large outside the scan, where the surrogate is unconstrained).

Surrogate accuracy and calibration



[20-fold cross-validation over all runs; empirical CDF of $|\tilde{f}_b - f_b|/\sigma_b$. Further left is better]

Calibration (held-out pulls $z = (\tilde{f} - f)/\sigma$):

$\sqrt{s}$ [TeV]	0.9	1.8	7	13
$\sqrt{\langle z^2 \rangle}$	1.013	1.052	0.966	1.006
$P(\lvert z \rvert < 2)$	94.8%	95.3%	96.1%	95.6%

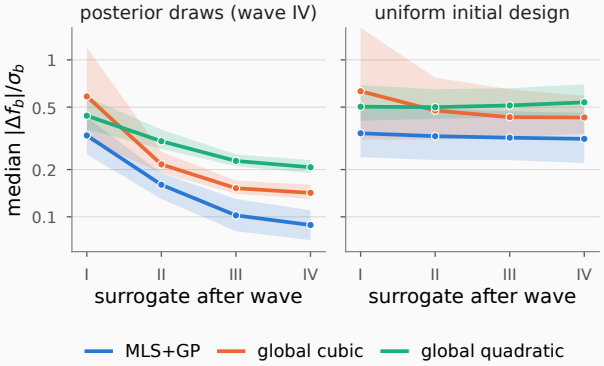
Accuracy: median |error| / data error

- **MLS+GP**: 0.13–0.17
- global cubic: 0.19–0.23
- global quadratic: 0.26–0.31

Even with the best surrogate (MLS+GP), surrogate error is $\sim$ 15% of data uncertainty

. . . but for MLS+GP, it is locally refinable

Refining the interpolation design, where it matters

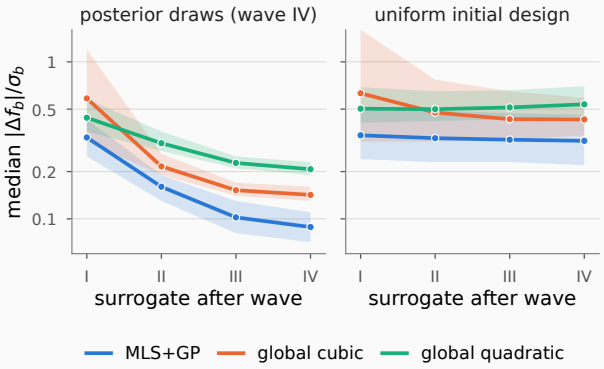


[held-out median error (line); range over the four energies (band)]

Refinement waves II–IV: new generator runs drawn from the current posterior, 319–910 → 1591–2341 per energy.

- *Near posterior*: MLS+GP error falls 4x; quadratic, cubic saturate
- *Elsewhere*: MLS+GP unchanged; global fits degrade

Refining the interpolation design, where it matters



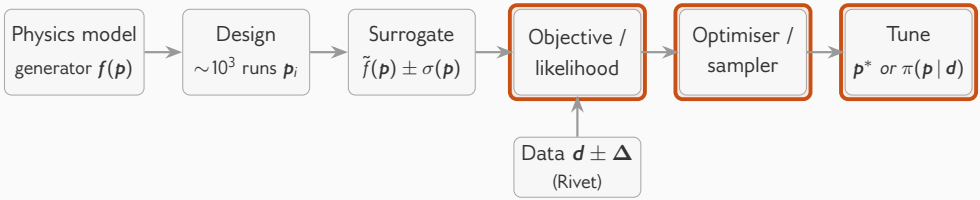
[held-out median error (line); range over the four energies (band)]

Refinement waves II–IV: new generator runs drawn from the current posterior, 319–910 → 1591–2341 per energy.

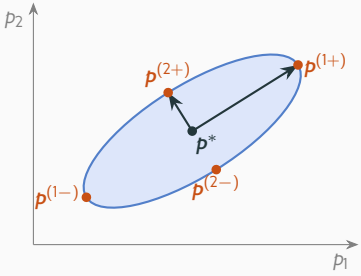
- *Near posterior*: MLS+GP error falls 4x; quadratic, cubic saturate
- *Elsewhere*: MLS+GP unchanged; global fits degrade

Iteration + validation gives a systematically-improveable method, as long as the surrogate can absorb new information

IV. Bayesian inference and MCMC



Classical tune uncertainties: ‘eigentunes’



$$\boldsymbol{p}^{(k\pm)} = \boldsymbol{p}^* \pm \sqrt{\lambda_k}\, \boldsymbol{t}_k$$

Local covariance $\mathbf{C}$ at the minimum $\Rightarrow$ $2d$ variation tunes along its eigenvectors.

- **Fisher** $\mathbf{F}^{-1}$: $\Delta\chi^2 = 1$ ellipse, based on assumption residuals are Gaussian;
- **Sandwich** $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\mathsf{T}}$: calculated from the observed residuals, so widens when the fit is poor.

Limitations

Local (uses second derivatives of surrogate), symmetric by construction, expansion only around a single minimum/mode; the definitions can differ by large factors; $\Delta\chi^2$ threshold is a convention.

Tuning as inference

$$\underbrace{\pi(\boldsymbol{p} \mid \boldsymbol{d})}_{\text{posterior}} = \frac{\overbrace{\pi(\boldsymbol{d} \mid \boldsymbol{p})}^{\text{likelihood}} \overbrace{\pi(\boldsymbol{p})}^{\text{prior}}}{\pi(\boldsymbol{d})}$$

- A distribution over parameters, not a point.
- Equally good choices of $\boldsymbol{p}$ (model 'equifinality') are shown

2 Predictions for *any* observable $\boldsymbol{F}$, fitted or not:

$$\pi(\boldsymbol{F} \mid \boldsymbol{d}) = \int \delta\big(\boldsymbol{F} - \boldsymbol{f}(\boldsymbol{p})\big)\, \pi(\boldsymbol{p} \mid \boldsymbol{d})\, \mathrm{d}\boldsymbol{p}$$

Draw $\boldsymbol{p}_i \sim \pi(\boldsymbol{p} \mid \boldsymbol{d})$, evaluate $\boldsymbol{f}(\boldsymbol{p}_i) \Rightarrow$ uncertainty bands.

Our posterior

Student-t likelihood, width $\sigma_b(\boldsymbol{p})$ including the surrogate error ($\nu \to \infty$: Gaussian):

$$\log \pi_t(\boldsymbol{d} \mid \boldsymbol{p}) = \sum_b \Big[-w_b^2 \, \frac{\nu+1}{2} \log\Big(1 + \frac{\big(\tilde{f}_b(\boldsymbol{p}) - d_b\big)^2}{\nu \, \sigma_b^2(\boldsymbol{p})}\Big) - \tfrac{1}{2} \log \sigma_b^2(\boldsymbol{p}) \Big]$$

Energies tempered, each counted as one mean bin:

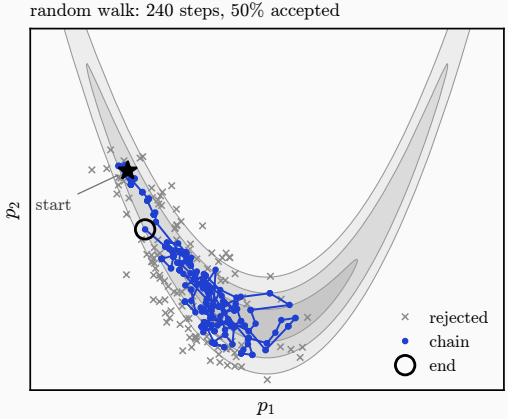
$$\log \pi(\boldsymbol{p} \mid \{\boldsymbol{d}_E\}) = \log \pi(\boldsymbol{p}) + \sum_E \frac{1}{n_E} \log \pi(\boldsymbol{d}_E \mid \boldsymbol{p})$$

Prior: broad Gaussian cores + smooth walls at the generator's limits (flat-prior test in VI).

A generalised posterior: energy-tempering sets the width, so should be calibrated

[Lange et al. (1989), Bissiri et al. (2016)]

Sampling the posterior: Markov chain Monte Carlo



Markov chain: built so that its stationary distribution is the posterior (10D grid would need 10^{10} points).

Metropolis–Hastings: accept $\boldsymbol{p}'$ with probability

$$\min\left\{1, \frac{\pi(\boldsymbol{p}'|\boldsymbol{d})}{\pi(\boldsymbol{p}|\boldsymbol{d})}\right\}$$

Random walk is *diffusive*: small steps, many rejections, correlated draws.

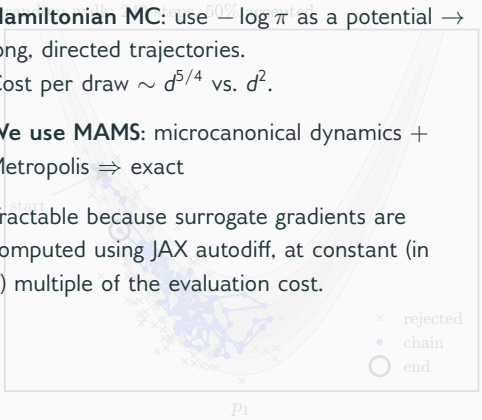
[Metropolis et al. (1953), Hastings (1970), Duane et al. (1987), Robnik et al. (2503.01707), Cabezas et al. (2402.10797)]

Sampling the posterior: Markov chain Monte Carlo

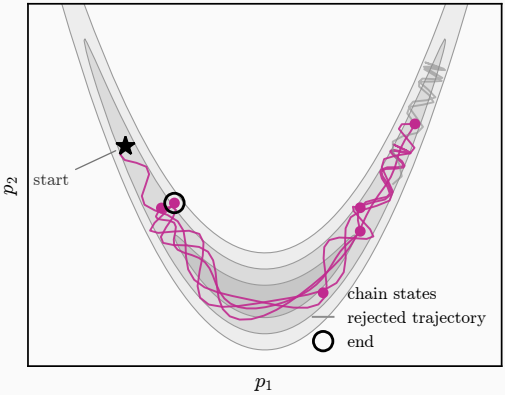
Hamiltonian MC: use $-\log \pi$ as a potential $\rightarrow$ long, directed trajectories.
Cost per draw $\sim d^{5/4}$ vs. d^2.

We use MAMS: microcanonical dynamics + Metropolis $\Rightarrow$ exact

Tractable because surrogate gradients are computed using JAX autodiff, at constant (in d) multiple of the evaluation cost.

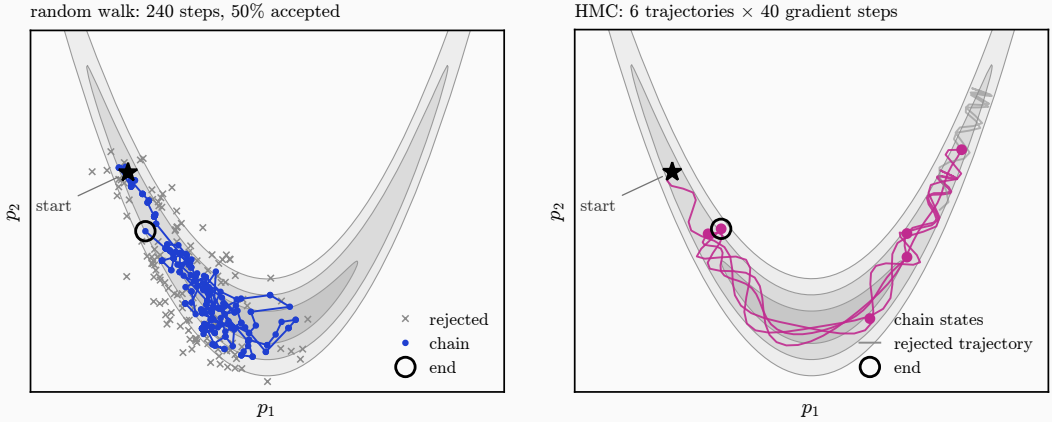


HMC: 6 trajectories × 40 gradient steps



[Metropolis et al. (1953), Hastings (1970), Duane et al. (1987), Robnik et al. (2503.01707), Cabezas et al. (2402.10797)]

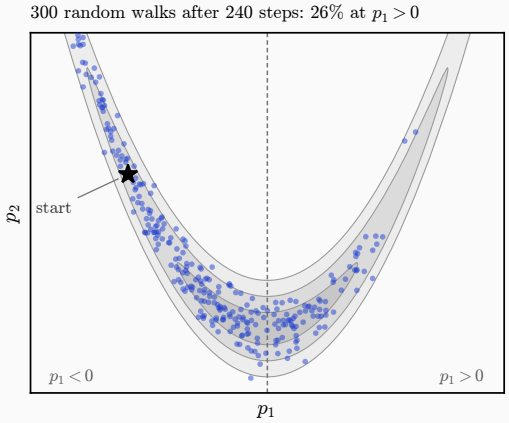
Sampling the posterior: Markov chain Monte Carlo

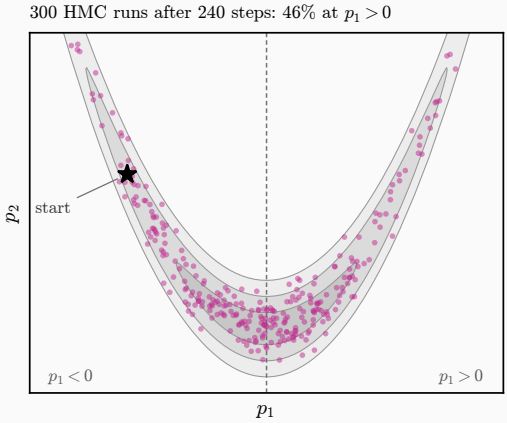


Same budget of 240 steps: the random walk diffuses, the gradient sampler explores.

[Metropolis et al. (1953), Hastings (1970), Duane et al. (1987), Robnik et al. (2503.01707), Cabezas et al. (2402.10797)]

Sampling the posterior: Markov chain Monte Carlo

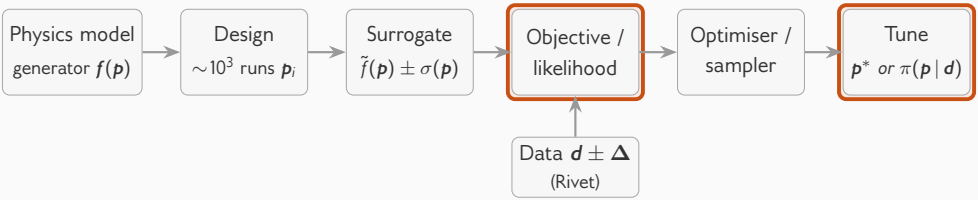




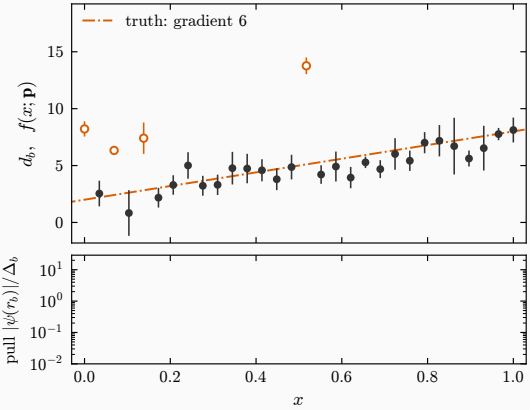
After 240 steps, 26% of random walks vs. 46% of HMC runs have reached $p_1 > 0$ (target: 50%).

[Metropolis et al. (1953), Hastings (1970), Duane et al. (1987), Robnik et al. (2503.01707), Cabezas et al. (2402.10797)]

V. Assembling the pieces: a toy example



A toy model where we know the truth



model $f(x;\boldsymbol{p}) = p_0 + p_1 x$, 30 bins; truth $\boldsymbol{p} = (2, 6)$

- **(a)** 15% of bins shifted up ('contamination')
- **(b)** a bump (model inadequacy)

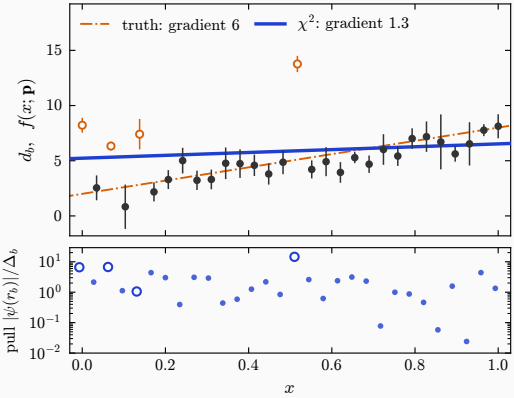
Pull of a gross outlier:

$$\frac{\psi(r_b)}{\Delta_b} \to \begin{cases} (d_b - f_b)/\Delta_b^2 & \chi^2 \\ \delta/\Delta_b & \text{Huber} \\ (\nu+1)/(d_b - f_b) & \text{Student-}t \end{cases}$$

Only a redescending loss makes the pull independent of Δ_b.

Two contamination scenarios

(a) random contamination



(b) structured discrepancy

15% of bins shifted up

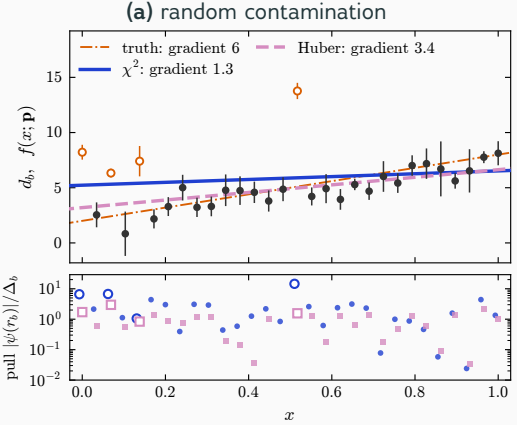
Gradient (truth 6), and contaminants' share of total pull:

	gradient	outlier pull
χ^2	1.3	40%

χ^2 follows the outliers.

[Kennedy and O'Hagan (2001), Brynjarsdóttir and O'Hagan (2014)]

Two contamination scenarios



15% of bins shifted up

Gradient (truth 6), and contaminants' share of total pull:

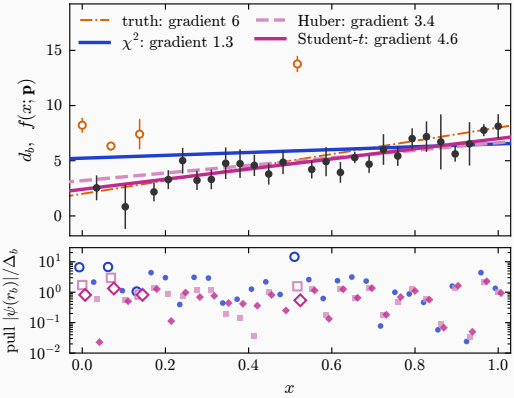
	gradient	outlier pull
χ^2	1.3	40%
Huber	3.4	26%

Huber bounds the pull, but a precise outlier still pulls $\propto 1/\Delta_b$.

[Kennedy and O'Hagan (2001), Brynjarsdóttir and O'Hagan (2014)]

Two contamination scenarios

(a) random contamination



(b) structured discrepancy

15% of bins shifted up

Gradient (truth 6), and contaminants' share of total pull:

	gradient	outlier pull
χ^2	1.3	40%
Huber	3.4	26%
Student-t	4.6	16%

Student-t is more *robust* under contamination

[Kennedy and O'Hagan (2001), Brynjarsdóttir and O'Hagan (2014)]

Two contamination scenarios

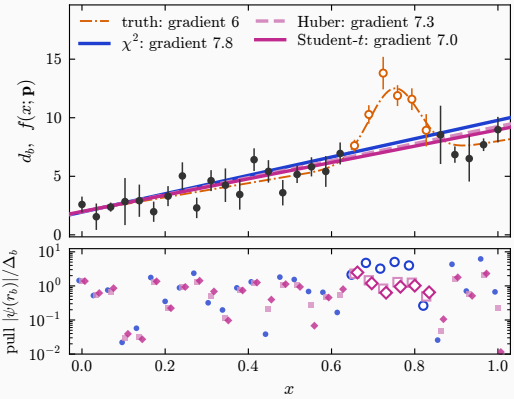
Bump (model inadequacy)

Peak down-weighted, but the shoulder bins look ordinary and still pull.

Tangential discrepancy: absorbed into $\boldsymbol{p}$, invisible per bin.

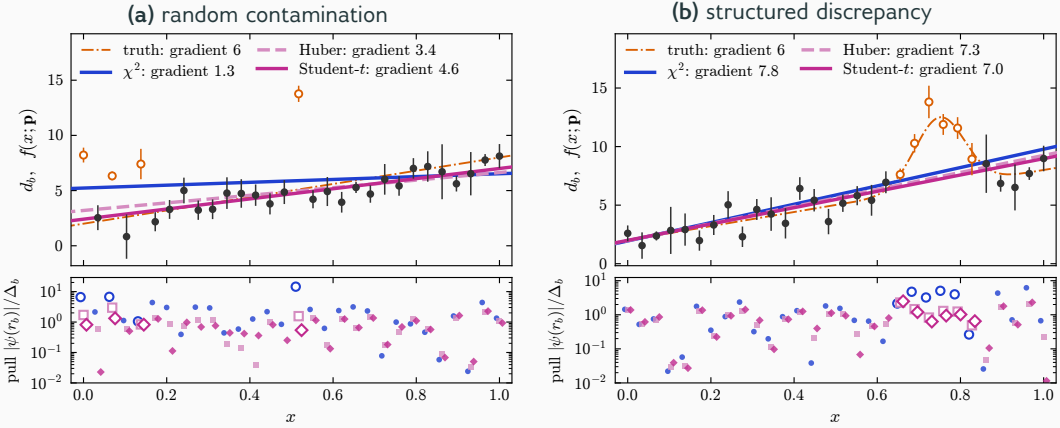
A more flexible model absorbs *more* ('identifiability').

(b) structured discrepancy



[Kennedy and O'Hagan (2001), Brynjarsdóttir and O'Hagan (2014)]

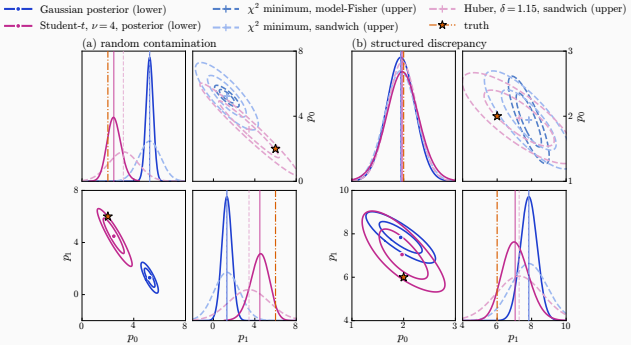
Two contamination scenarios



Robustness can limit the damage done by large residuals — but cannot fix model inadequacy

[Kennedy and O'Hagan (2001), Brynjarsdóttir and O'Hagan (2014)]

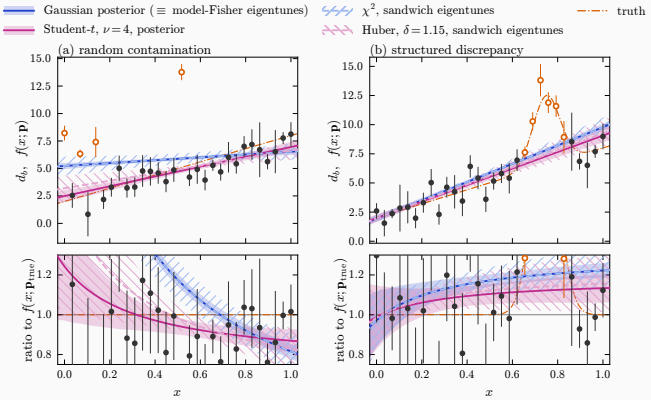
Uncertainties on the toy



Eg: fit of (a), $\chi^2/\mathrm{ndf} = 8$:

- Gaussian: truth 11.8σ away
- Inflated by $\sqrt{\chi^2/\mathrm{ndf}}$, or sandwich: wider, *not moved*
- Huber, Student-t: wider *and* moved; both cover the truth

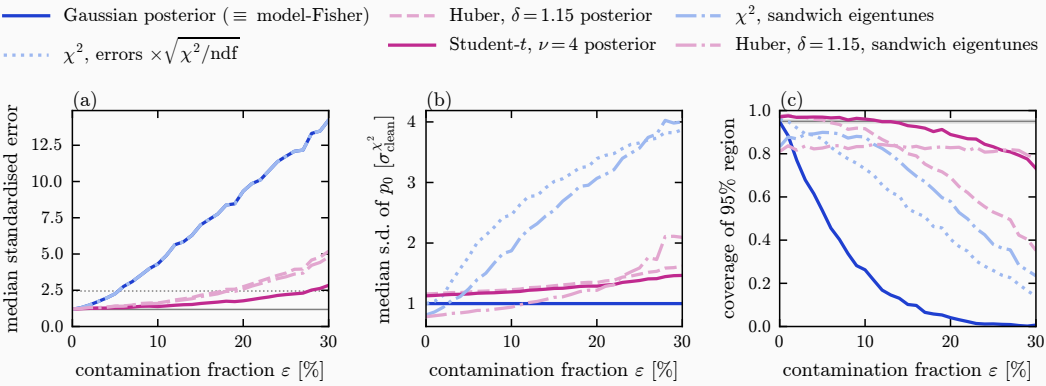
Uncertainties on the toy



Posterior-predictive $\pm 1\sigma$ bands for $f(x)$

- (a) χ^2 sandwich band is $\sim 3\times$ wider, but still centred on the displaced line
- (b) structured discrepancy still partially absorbed

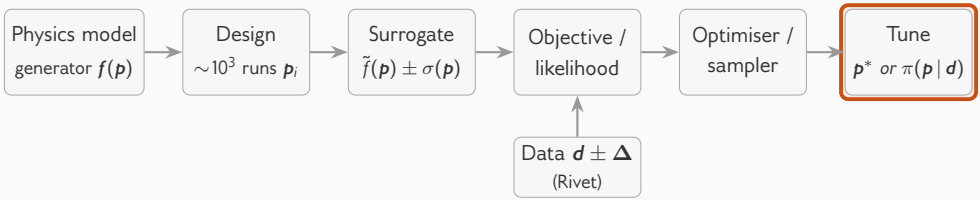
Uncertainties on the toy



[Same fit, 2000 repetitions]

price at $\varepsilon = 0$: wider intervals.
in tuning, ε is unknown, but almost certainly $\neq 0$

VI. Results: re-tuning LH stage 4



Our test case: LH-tune stage 4, redone

The LH tune

Herwig 7 shower + Pythia 8 string and CR

1. **string fragmentation, FSR**
   LEP: event shapes, spectra
   ↓ freeze
2. **flavour**
   LEP: hadron yields
   ↓
3. **ISR α_s, intrinsic $k_\perp$**
   Drell–Yan, 7 TeV
   ↓
4. **MPI, CR**: 10 parameters — redone here
   MB & UE, 0.9–13 TeV: 3589 bins

Pythia 8.311+ changed string and CR code so LH stage 4 is not reproducible.
Idea: repeat with same data and weights.

	parameter	initial range[†]
MPI / UE (Herwig)	c (energy exponent)	$[0, 2]$
	$p_{\perp,0}^{\min}$	$[2, 5]$ GeV
	b (energy offset)	$[500, 800]$ GeV
	μ^2 (overlap)	$[0.5, 3.5]$ GeV^2
	N_{ladder}	$[0.5, 2]$
	B_{ladder}	$[0.5, 2]$
	R_{diff}	$[0, 0.5]$
CR	m_0	$[0.1, 3]$ GeV
	m_{pseudo} *(new)*	$[0.1, 3]$ GeV
	`junctionCorrection`	$[0.05, 5]$

[†] widened by the refinement waves (slide 23).

[Divisova et al. (2509.02348), Bierlich et al. (2203.11601), Pythia 8 update history]

What we compare

Bayesian: Gaussian; Student-t, $\nu = 4$ (also 3, 6).

Classical: χ^2; Huber, $\delta = 1.15$ (also 1, 1.345); sandwich eigentunes.

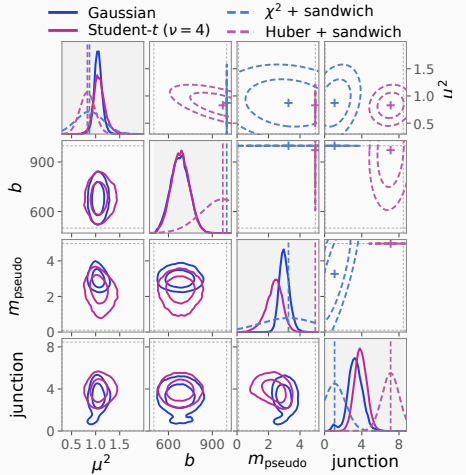
Variations

- robustness: ν, δ;
- analysis weights: $w_b = 5 \to 10, 3, 1$;
- per energy vs. per histogram;
- nominal vs. flat prior.

All on the same surrogate, data, weights.
Reference: the LH tune (same physics model, older Pythia code)

preliminary: preprint expected Nov 2026

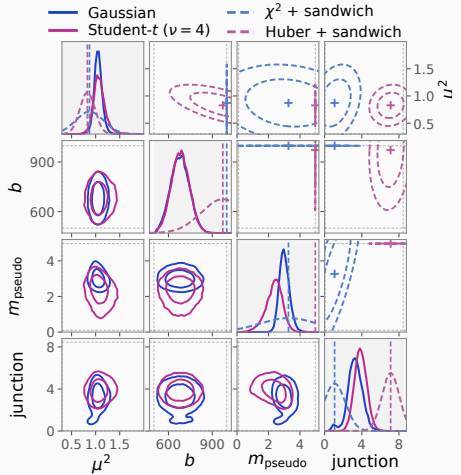
Headline: posteriors vs. classical minima



[4 of 10 parameters; μ^2 in GeV^2, b, m_{pseudo} in GeV; dashed grey: training box]

Lower: Gaussian, Student-t posteriors.
Upper: χ^2, Huber minima + sandwich.

Headline: posteriors vs. classical minima



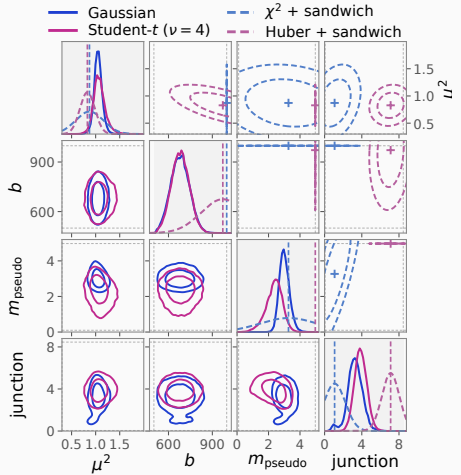
[4 of 10 parameters; μ^2 in GeV^2, b, m_{pseudo} in GeV; dashed grey: training box]

Lower: Gaussian, Student-t posteriors.

Upper: χ^2, Huber minima + sandwich.

- posteriors agree (largest shift 1.3σ)

Headline: posteriors vs. classical minima



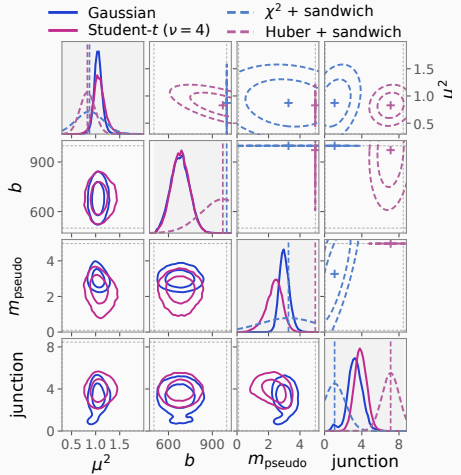
[4 of 10 parameters; μ^2 in GeV^2, b, m_{pseudo} in GeV; dashed grey: training box]

Lower: Gaussian, Student-t posteriors.

Upper: χ^2, Huber minima + sandwich.

- posteriors agree (largest shift 1.3σ)
- classical minima: different basins, either side of the posterior
- b at its upper limit, where the posterior is just the prior

Headline: posteriors vs. classical minima



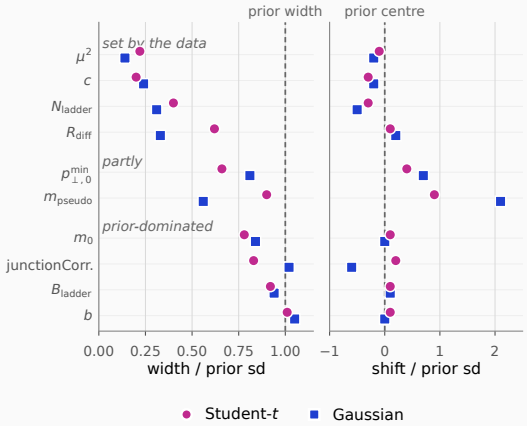
[4 of 10 parameters; μ^2 in GeV^2, b, m_{pseudo} in GeV; dashed grey: training box]

Lower: Gaussian, Student-t posteriors.

Upper: χ^2, Huber minima + sandwich.

- posteriors agree (largest shift 1.3σ)
- classical minima: different basins, either side of the posterior
- b at its upper limit, where the posterior is just the prior
- sandwich ellipses much wider in m_0, m_{pseudo}, B_{ladder}

Which parameters do the data constrain?

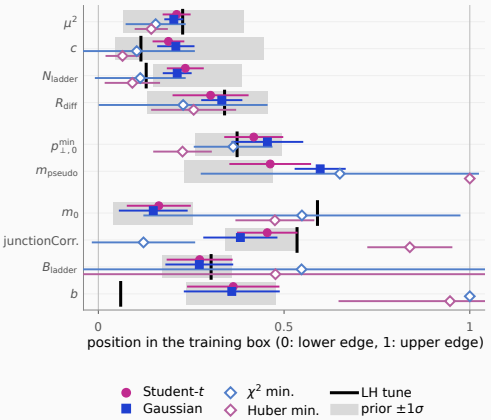


Each marginal vs. its prior: width and shift.

- **Set by the data**: μ^2, c, N_{ladder}, R_{diff}.
- **Not determined**: b, B_{ladder}, m_0, junctionCorr: the marginals are the prior.
- m_{pseudo}: pulled up by the $w_b = 5$ bins.

4 / 10 parameters are not constrained by these data: can inform data selection, model development

Compared with the published LH tune

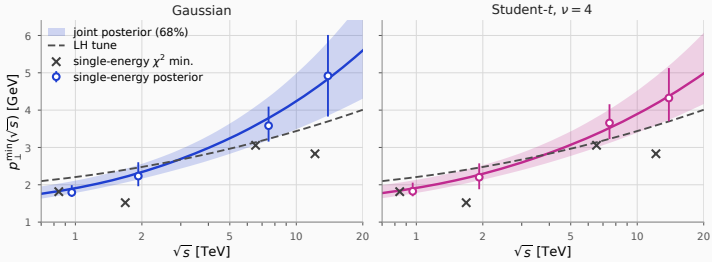


Rows ordered by data constraints (from last slide)

- **Set by the data**: μ^2, R_{diff}, $p^{\text{min}}_{\perp,0}$ agree with LH; c and N_{ladder} moved by $\sim 2\sigma$ (could be effect of changed code)
- m_{pseudo} is new: split from m_0; no LH comparator
- classical minima scatter across the box, several at its edges

Extent of agreement with LH tune is explained by data constraints

The MPI cutoff across the energies



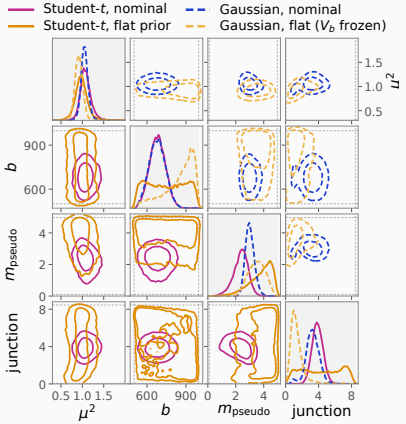
$p_\perp^{\min}(\sqrt{s}) = p_{\perp,0}^{\min}\,((b+\sqrt{s})/7\,\mathrm{TeV})^c$; bands and bars: 68%

Single-energy posteriors (same prior) agree with joint band: b in isolation is prior-dominated, but $p_\perp^{\min}(\sqrt{s})$ in combination is constrained

(Single-energy χ^2 minima are not consistent)

The data at each energy constrains the MPI cutoff consistently (but not with χ^2)

Remove the prior

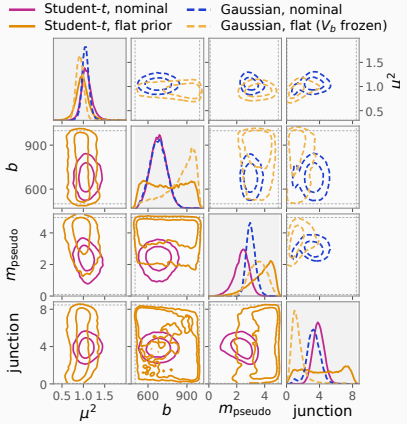


[flat prior: uniform over the training box, same walls]

Flat prior over the training box:

- Student-t: b, junctionCorr, m_0, $B_{\rm ladder}$ fill the box; μ^2, c, $N_{\rm ladder}$, $R_{\rm diff}$ barely move.

Remove the prior



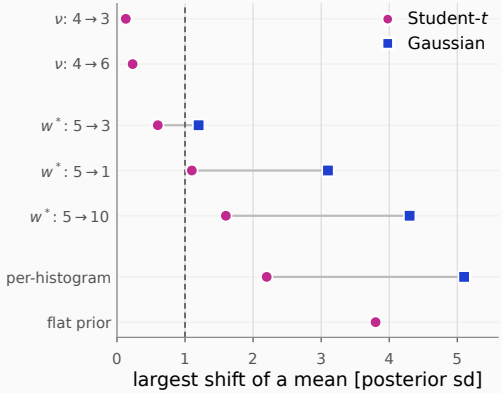
[flat prior: uniform over the training box, same walls]

Flat prior over the training box:

- Student-t: b, junctionCorr, m_0, B_{ladder} fill the box; μ^2, c, N_{ladder}, R_{diff} barely move.
- Gaussian: drifts to where the surrogate variance is large; with V_b frozen, it lands near the χ^2 minimum.

The prior, not the data, determines four fitted parameters – despite 3589 constraints

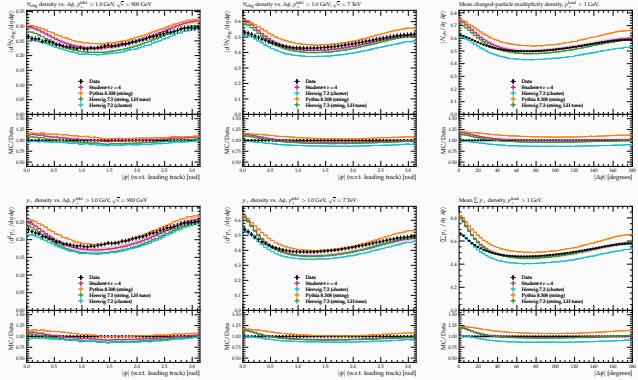
Sensitivity at a glance



- $\nu = 3, 4, 6$: indistinguishable.
- Weights: Student-t 0.6–1.6σ; Gaussian 1.2–4.3σ, changing basin.
- Per histogram: 3$\times$ narrower, and the likelihoods part ways.
- Flat prior: the largest shifts.

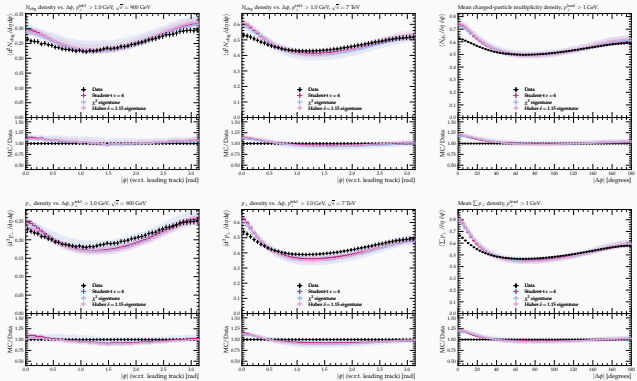
Gaussian $\rightarrow$ robust matters more than ν; robust posterior depends less on weights

Back to observables



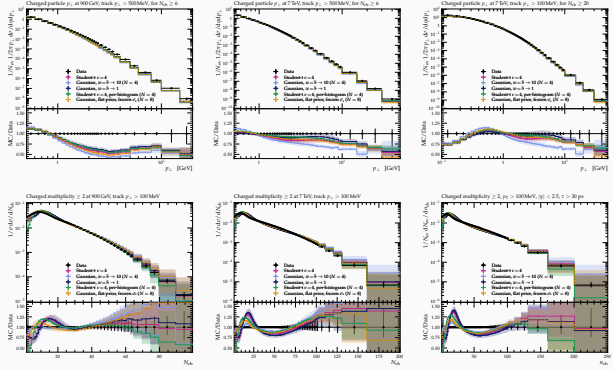
LH-tune Fig. 8 selection: N_{ch} (top) and $\sum p_{\perp}$ (bottom) vs. $\Delta\phi$ to the leading track. Student-t prediction, 68% band.

Back to observables



Student-t vs. χ^2, Huber: almost the same central predictions; all overshoot the toward region ($\Delta\phi \approx 0$) by 15–20% at 7, 13 TeV.

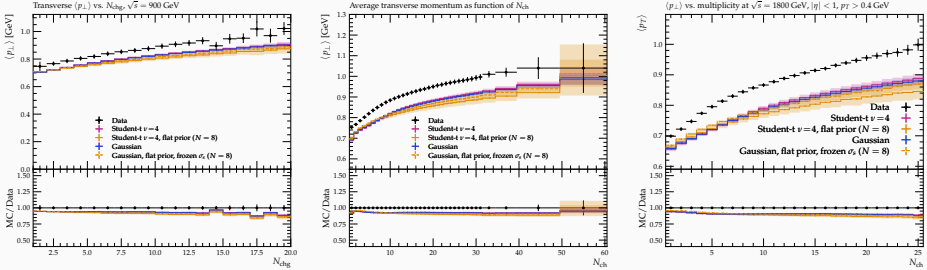
Trade-offs between distributions



Top: $p_\perp$ spectra; bottom: N_{ch}.
Student-t; Gaussian, $w^* = 10$ and 1; per histogram; flat prior.
w^*: $10 \to 1$ changes $\sum_b r_b^2/n$ by -30.4 to -12.7 (spectra) and $+4.9$ to $+11.9$ (multiplicities).

Changing the target trades one family of distributions against another.

Prior sensitivity: $\langle p_\perp \rangle(N_{\mathrm{ch}})$

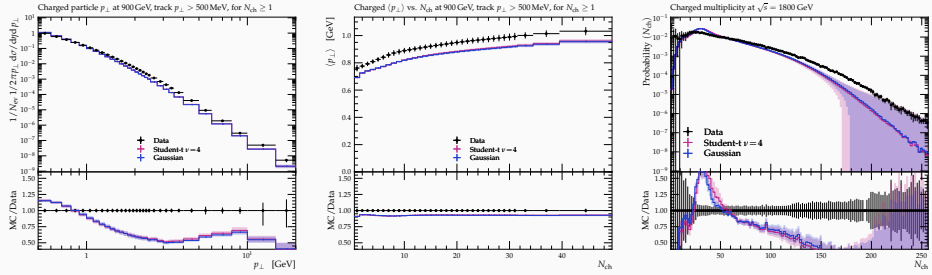


0.9 TeV transverse region (ATLAS) 0.9 TeV (ATLAS) 1.8 TeV (CDF)

- Nominal vs. flat prior: the Student-t predictions move by 0.9 of the 68% half-width (median histogram), but by 1.7–2.7 in $\langle p_\perp \rangle(N_{\mathrm{ch}})$ at 0.9 and 1.8 TeV.
- In all 6 such distributions $\langle p_\perp \rangle$ drops further below the data: these are the most CR-sensitive observables, and the CR parameters are undetermined.

Prior sensitivity highlights the effect of unconstrained parameters.

Model inadequacy

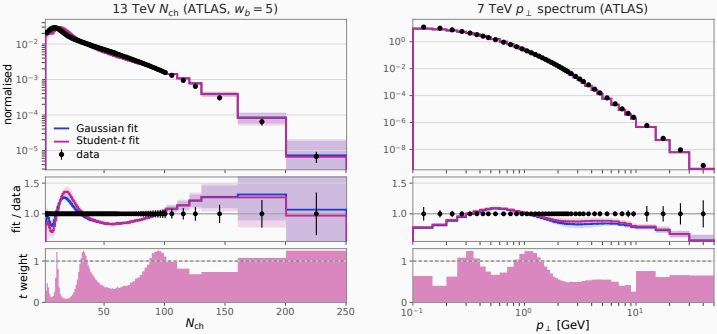


0.9 TeV charged $p_\perp$ spectrum 0.9 TeV $\langle p_\perp \rangle$ vs. N_{ch} 1.8 TeV N_{ch} (E735)

- 0.9 TeV $p_\perp$ spectra: 63% of bins beyond 2σ; 5–8% low, in every bin (NB: systematic uncertainty correlations neglected in likelihood).
- Consistently across likelihoods, weights, priors: the model is too soft at low energy.

Robust inference shows where the model fails, but cannot fix it.

Model inadequacy



Robust pulls in practice: bins the model cannot describe get a small Student-t weight $(\nu + 1)/(\nu + r_b^2)$ (bottom).
The 11% of bins with $|r_b| > 3$ carry 32% of the Gaussian fit's pull, but only 13% of the Student-t's.

The Gaussian fit prioritises a small improvement in extreme outliers above many improvements across many bins

Summary

1. **Central prediction**: outliers are *structural*; a Student-t likelihood is less sensitive to the weight choice, and can find the right region even with a crude surrogate.
2. **Uncertainties for any bin**: posterior samples from gradient-based MCMC on a calibrated surrogate, in hours on a GPU; usable like replica tunes.
3. **Model inadequacy**: 4 of 10 parameters undetermined, in part because the energies disagree; low-energy spectra too soft.

χ^2 χ^2 **alone**: minima in one of two basins, error bars differing by $\times$2.2–4.7; some at box limits.

Robustness limits the damage of model inadequacy; it does not remove it.

Outlook

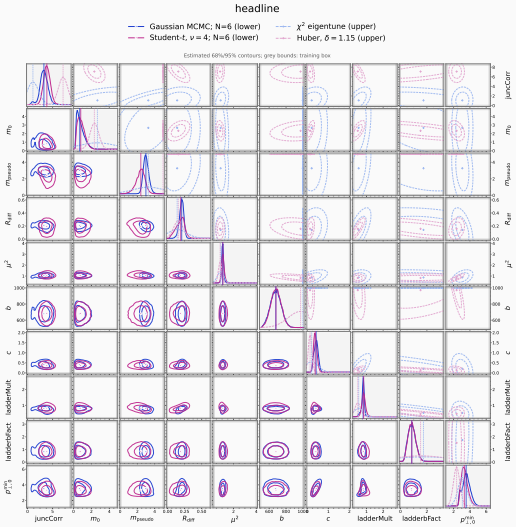
- **Refinement on autopilot**: iteration towards a convergence criterion instead of hand-placed waves
- **Calibrate** the tempered posterior with closure tests
- **Energy tension**: a modelling question (energy dependence of MPI and CR)
- **Correlated data**: *want* full covariances, e.g. DELPHI open data (backup)
- **Earlier stages, several models**, chained through priors (posterior of one stage is the prior of the next)
- **For experiments**: propagable tune uncertainty as posterior samples
- **For Herwig**: possibility of simultaneous, consistent, comparable, dipole/angular-ordered cluster/string tunes.

Thank you.

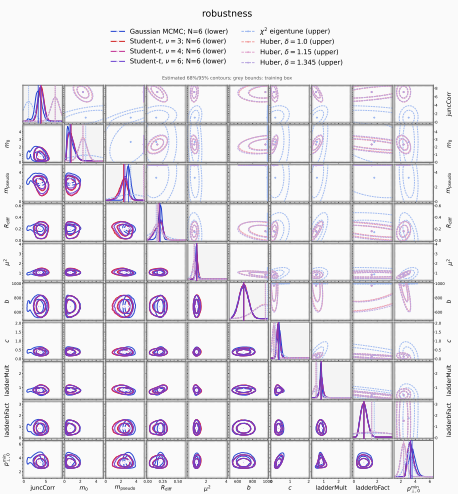
Questions?

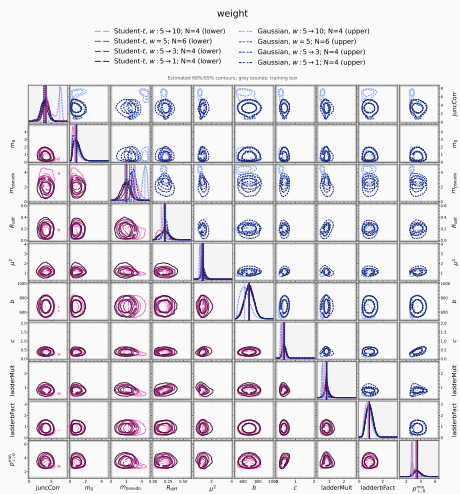
Backup

Backup: full headline corner (all 10 parameters)

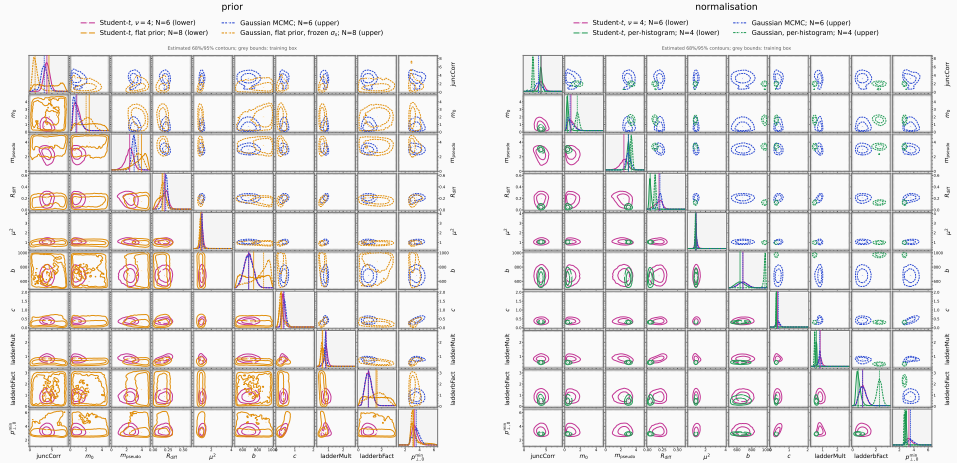


Backup: robustness and weight ladders (all parameters)



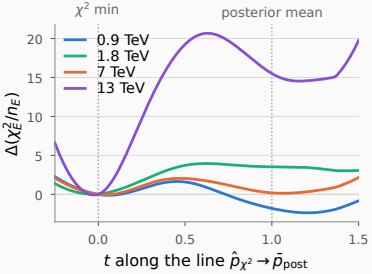


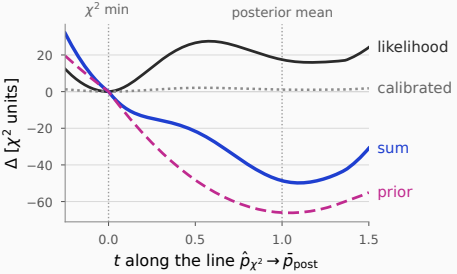
Backup: prior and normalisation (all parameters)



Left: nominal vs. flat prior (Student-t lower; Gaussian, V_b frozen, upper). Right: per energy ($1/n_E$) vs. per histogram normalisation.

Backup: cause of prior dependence: the energies disagree





- χ^2 minimum ($t = 0$) $\to$ posterior mean ($t = 1$): $\Delta(\chi^2_E/n_E) = +15.5$ at 13 TeV, $+3.5$ at 1.8 TeV, -1.8 at 0.9 TeV.
- Tempered likelihood $+17.4$ vs. prior -66.1 (χ^2 units): the prior arbitrates.

Different energies prefer different individual MPI and CR parameters

Backup: proof of Huber's minimax theorem

Theorem (Huber 1964)

Let residuals $r \sim F \in \mathcal{F}_\varepsilon = \{(1-\varepsilon)\Phi + \varepsilon H\}$, with H symmetric but otherwise arbitrary. Among all M-estimators ($\sum_b \psi(r_b) = 0$, asymptotic variance $V(\psi, F) = \mathbb{E}_F[\psi^2]/(\mathbb{E}_F[\psi'])^2$), the worst case $\sup_{F\in\mathcal{F}_\varepsilon} V(\psi, F)$ is minimised by $\psi_{\mathrm{H}} = \max(-\delta, \min(r, \delta))$, with $\frac{2\phi(\delta)}{\delta} - 2\Phi(-\delta) = \frac{\varepsilon}{1-\varepsilon}$.

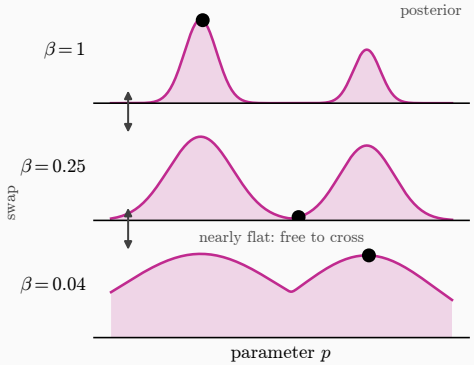
Proof: a saddle point. Exhibit $F_0 \in \mathcal{F}_\varepsilon$ with

$$V(\psi_{\mathrm{H}}, F) \overset{\text{(b)}}{\leqslant} V(\psi_{\mathrm{H}}, F_0) \overset{\text{(a)}}{\leqslant} V(\psi, F_0) \quad \forall\, F \in \mathcal{F}_\varepsilon,\ \forall\, \psi \qquad \Longrightarrow \qquad \sup_F V(\psi_{\mathrm{H}}, F) \leqslant V(\psi, F_0) \leqslant \sup_F V(\psi, F).$$

1. **Guess** $f_0 \propto e^{-\rho_{\mathrm{H}}}$: Gaussian core, exponential tails, with $f_0 = (1-\varepsilon)\phi$ for $|r| \leqslant \delta$. Then $\int f_0 = 1$ *is* the δ–ε relation, and the excess $f_0 - (1-\varepsilon)\phi \geqslant 0$ lives on $|r| > \delta$, so $F_0 \in \mathcal{F}_\varepsilon$.
2. **(a), Cramér–Rao for M-estimators**: by parts, $\mathbb{E}_F\psi' = \mathbb{E}_F[\psi\,\ell]$ with score $\ell = -f'/f$; Cauchy–Schwarz gives $V(\psi, F) \geqslant 1/I(F)$, with equality iff $\psi \propto \ell$. The score of f_0 is exactly ψ_{H}: Huber is the MLE for F_0.
3. **(b), worst case**: $|\psi_{\mathrm{H}}| \leqslant \delta$ and $\psi'_{\mathrm{H}} \geqslant 0$ give $\mathbb{E}_F\psi^2_{\mathrm{H}} \leqslant (1-\varepsilon)\mathbb{E}_\Phi\psi^2_{\mathrm{H}} + \varepsilon\delta^2$ and $\mathbb{E}_F\psi'_{\mathrm{H}} \geqslant (1-\varepsilon)\mathbb{E}_\Phi\psi'_{\mathrm{H}}$. Both are equalities iff H lives on $|r| > \delta$, where ψ_{H} is saturated, as for F_0. □

[Huber (1964), Huber and Ronchetti (2009, ch. 4), Maronna et al. (2019)]

Backup: several modes, parallel tempering



- Gradient samplers rarely cross between separated modes
- **Parallel tempering**: replicas of $\pi(\boldsymbol{p})\,\pi(\boldsymbol{d} \mid \boldsymbol{p})^{\beta}$, $\beta = 1 \to 0.01$; hot replicas cross barriers, neighbours swap; keep $\beta = 1$
- In practice: the single-energy posteriors are *bimodal*

[Swendsen and Wang (1986), Geyer (1991), Earl and Deem (2005)]

Backup: the MLS + GP surrogate

$$\tilde{f}_b(\boldsymbol{p}) = \tilde{f}_b^{\,\mathrm{MLS}}(\boldsymbol{p}) + \tilde{f}_b^{\,\mathrm{GP}}(\boldsymbol{p})$$

MLS: local quadratic, coefficients refitted at each $\boldsymbol{p}$:

$$\boldsymbol{\alpha}^*(\boldsymbol{p}) = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_i \omega_i(\boldsymbol{p}) \big[\boldsymbol{\alpha} \cdot \boldsymbol{\pi}(\boldsymbol{p}_i) - f_b(\boldsymbol{p}_i)\big]^2,$$

$$\omega_i(\boldsymbol{p}) = \exp\Big[-\tfrac{1}{2}\sum_j \tfrac{(p_j - p_{ij})^2}{h_j^2}\Big]$$

GP: Matérn-$\frac{5}{2}$ kernel with per-parameter lengthscales, trained on MLS leave-one-out residuals; hyperparameters shared per energy.

Into the likelihood

Per-bin width, including the surrogate variance V_b:

$$\sigma_b^2(\boldsymbol{p}) = \max\left\{\Delta_b^2 + V_b(\boldsymbol{p}),\ (0.05\, d_b)^2\right\}$$

$V_b = V_b^{\mathrm{GP}}$ inside the scanned box; the MLS extrapolation variance is switched on smoothly outside it.

- the width now depends on $\boldsymbol{p}$;
- the 5% floor acts as a minimal model-discrepancy term.

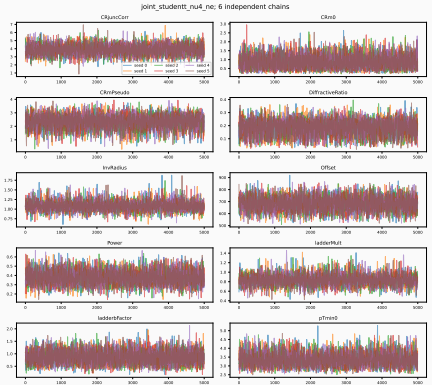
[Lancaster and Salkauskas (1981), Rasmussen and Williams (2006)]

Backup: stage IV data and weights

$\sqrt{s}$	Rivet analyses	weights
0.9 TeV	ALICE_2010_I860416; ATLAS_2011_I891834; ATLAS_2010_I882098; ATLAS_2010_I849050; ATLAS_2010_I879407	$w_b = 5$: I882098 η distributions at $(p_\perp, N_{\mathrm{ch}})$ thresholds $(100\,\mathrm{MeV}, 2)$, $(500\,\mathrm{MeV}, 1)$, $(500\,\mathrm{MeV}, 6)$, and I849050 $\mathrm{d}N_{\mathrm{ch}}/\mathrm{d}\eta$; else 1
1.8 TeV	CDF_2002_I567774; E735_1998_I480349; CDF_1990_I283352; CDF_1988_I263320; CDF_2001_I564673	all 1
7 TeV	ALICE_2010_I852264; ATLAS_2011_I891834; ATLAS_2010_I882098; ATLAS_2010_I879407	$w_b = 5$: the same three I882098 η selections; else 1
13 TeV	ATLAS_2017_I1509919; ATLAS_2016_I1467230	$w_b = 5$: all four I1467230 distributions; 1 for I1509919, with $15 < p_\perp^{\mathrm{lead}} < 30$ GeV bins at 0

[Effective weighting per energy, $A_E = n_E^{-1} \sum_{b \in E} w_b^2$: 4.98, 1.00, 3.62, 11.49 at 0.9, 1.8, 7, 13 TeV.
As a result, one 13 TeV bin pulls $\sim 11\times$ harder than one 1.8 TeV bin.]

Backup: in practice (sampler, diagnostics, cost)

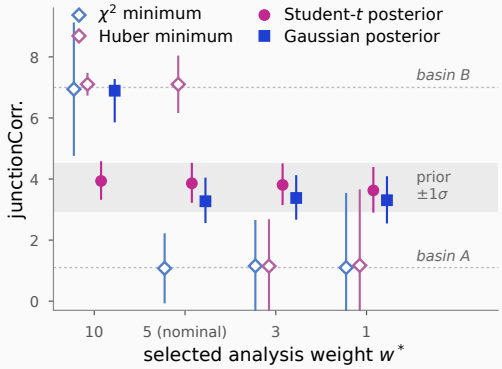


[all 10 parameters, 6 independent chains, 5×10^3 draws each]

- PT-MAMS in BlackJAX: 24 replicas, $\beta_{\min} = 10^{-2}$, 10^4 warm-up + 5×10^3 draws per replica, seeded at the best of 100 L-BFGS-B optimisations.
- 6 independent chains per nominal fit: $\widehat{R} \leq 1.002$; bulk and tail ESS $\geqslant 2800$. All 15 fit variants: $\widehat{R} < 1.01$.
- Cost per chain: $\sim$1 day on a laptop (6 cores), or 4–6 h on one V100 GPU; 4 chains vectorised on one GPU in 6–12 h.
- **Classical baseline** on the *same* surrogate: multi-start L-BFGS-B, χ^2 and Huber, with sandwich eigentunes (also in JAX).

[Vehtari et al. (1903.08008), Byrd et al. (1995)]

Backup: a tale of two basins

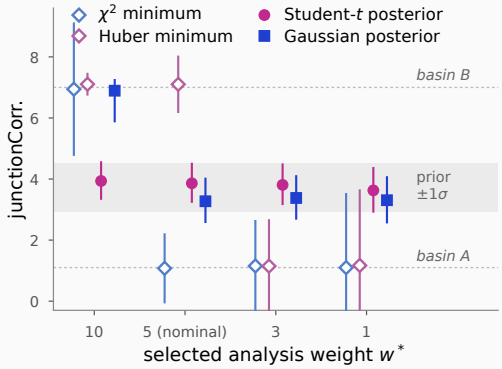


[bars: ± 1 sandwich sd (classical), 68% interval (posterior)]

100 fits end in **basin A** (junctionCorr ≈ 1) or **B** (≈ 7, m_{pseudo} at its box edge).

- χ^2 prefers A; Huber B, narrowly (3/100 starts find it).

Backup: a tale of two basins

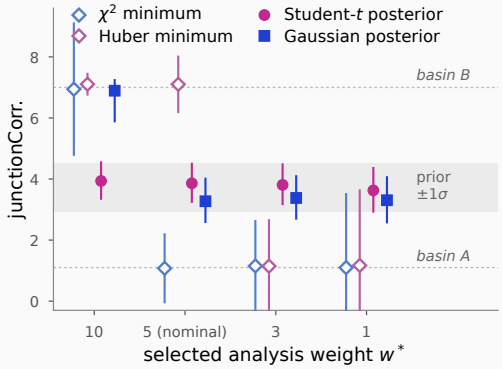


[bars: ±1 sandwich sd (classical), 68% interval (posterior)]

100 fits end in **basin A** (junctionCorr ≈ 1) or **B** (≈ 7, m_{pseudo} at its box edge).

- χ^2 prefers A; Huber B, narrowly (3/100 starts find it).
- Change the weights and both jump: pushed from one minimum to another by the changed trade-off between distributions

Backup: a tale of two basins



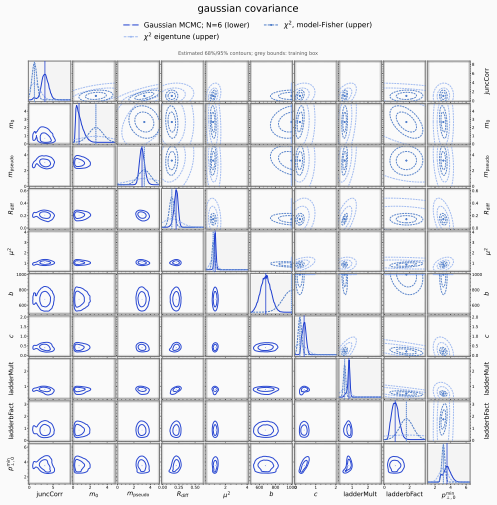
[bars: ± 1 sandwich sd (classical), 68% interval (posterior)]

100 fits end in **basin A** (junctionCorr ≈ 1) or **B** (≈ 7, m_{pseudo} at its box edge).

- χ^2 prefers A; Huber B, narrowly (3/100 starts find it).
- Change the weights and both jump: pushed from one minimum to another by the changed trade-off between distributions
- Student-t: stays between, near its prior.

A point estimate plus local error bars cannot represent this.

Backup: Fisher vs sandwich at the χ^2 minimum



At the *same* χ^2 minimum:

- Fisher $\mathbf{F}^{-1}$: assumes the residuals are unit Gaussians;
- empirical sandwich: uses the observed residuals.

Sandwich widths of the free parameters are $\times$2-5 the Fisher ones; for m_0, $m_{\rm pseudo}$, $B_{\rm ladder}$ wider than the prior.

Both are solid ‘textbook’ choices; the posterior width is instead governed by the $1/n_E$ tempering.