LPCC MCWG Tuning Forum

CMS tuning

Experience with Professor, jet-substructure methodology and new tools

Bora Işıldak, Eda Nur Edoğan
Deniz Sunar Çerçi, Salim Çerçi
Yildiz Technical University
for the CMS Collaboration

22 September 2026

Scope of this update

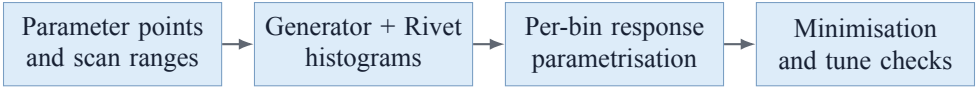
1. **Professor experience in CMS**
   Practical limitations and feedback from tuning campaigns.

2. **Jet-substructure tuning methodology**
   Lessons from surrogate tests and GP-guided sampling.

3. **MCNNTUNES evaluation**
   Public model modes and a plan for a like-for-like CMS test.

CMS tune overview by G. Van Onsem, March 2026 Tuning Forum.

The Professor tuning workflow



Why the approach is useful

- Parallel generator production.
- Fast repeated fits and scans.
- Reuse with different observable weights.

Experience in the LJP study

Professor2 provides a common final interpolator while we test how the generator points are chosen.

The selected tune still needs a fresh generator run.

Professor method, Buckley et al., arXiv 0907.2973. LJP example from the CMS Generator Meeting, July 2026 (internal).

Professor2 limitations and validation checks

Response approximation

Each bin gets its own polynomial of order 2 in the d tuned parameters, with

$$n_{\text{coeff}} = \binom{d+2}{2}$$

Six parameters give 28 coefficients per bin and eleven give 78.

- Professor2 fits only the supplied MC points. A separate method must choose new points.
- Parameter range, polynomial order and sparse bins affect prediction accuracy.

Objective and uncertainty

- Training-MC errors enter through a fixed per-bin scale, not a model uncertainty.
- The default score ignores bin-to-bin correlations.
- The score guides optimisation. It is not a formal χ^2 probability.
- The standard workflow has no constrained optimiser.

Professor method, Buckley et al., Eur. Phys. J. C 65 (2010) 331.

Validation. Test the response on withheld MC points, then confirm the selected tune with new generator runs.

CMS experience and questions for users

Lessons from our LJP campaign

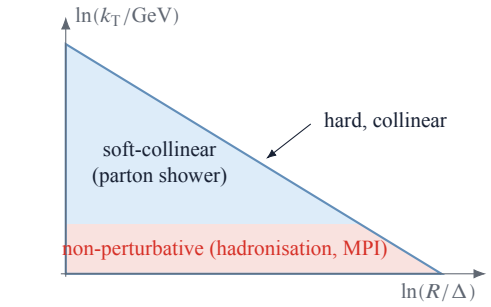
- A low predicted score can shift after full-generator validation.
- Tail bins require explicit treatment of MC statistical errors.
- Sampling changes can improve the same Professor2 fit.

Questions for wider CMS input

- Campaign size and software version.
- Diagnostics, minimisation and uncertainty tools.
- Reproducible failures and successful workarounds.
- Requested improvements, with examples.

LJP lessons from the CMS Generator Meeting, July 2026 (internal). The left column is our own campaign experience. The CMS Professor user questionnaire is still to be circulated. Findings at a future forum.

Jet-substructure tuning setup



One measurement constrains shower and hadronisation together.

Six Pythia8 parameters in this study

Shower	α_s^{FSR}, $p_{\mathrm{T,min}}^{\mathrm{FSR}}$
	α_s^{ISR} (under study)
Hadronisation	σ_{p_T}, a_{Lund}, b_{Lund}

$N = 50$ to 200 training runs, with 10,000 events per training point.

Fixed parameter box and active-bin definition for both sampling branches.

Active bins 137 (AK4), 138 (AK8), 275 combined.

The CP5 and CH3 tunes do not describe jet-substructure observables well. The Lund plane shows in which region the disagreement sits.

Measurement by the CMS Collaboration, JHEP 05 (2024) 116 (CMS-SMP-22-007), with an internal Rivet routine. Method study from the CMS Generator Meeting, July 2026 (internal).

Tune selection and generator validation

Professor2 selection

$$S_{\text{data}}(\boldsymbol{\theta}) = \frac{1}{N_{\text{active}}} \sum_{b \in \mathcal{B}} \frac{(\widehat{H}_b(\boldsymbol{\theta}) - D_b)^2}{\sigma^2_{\text{data},b}}$$

Quadratic Professor2. One fit uses all training points, and four additional fits each omit one group of points.

Choose the parameter point with the lowest 80th-percentile score across the five fits.

Full-generator validation

$$S_{\text{data+MC}} = \frac{1}{N_{\text{active}}} \sum_{b \in \mathcal{B}} \frac{(H_b - D_b)^2}{\sigma^2_{\text{data},b} + \sigma^2_{\text{MC},b}}$$

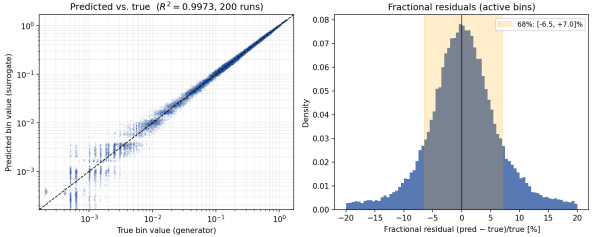
Five independent runs of 100,000 events at each selected point.

Report the mean and sample spread of the per-seed scores.

Both use the same 275 bins. These diagonal scores have different denominators.

Selection and validation conventions from the CMS Generator Meeting, July 2026 (internal). Here N_{active} counts bins, without subtracting fit parameters.

An early validation lesson from alternative surrogates



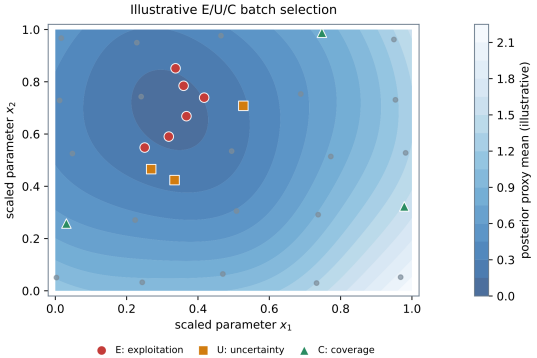
Method	Predicted median	Predicted Q80	First MC validation
Global DNN surrogate	2.342	2.632	**4.013**
Professor2, order 2	2.585	2.618	**2.927**

Same 275 bins and data-only score in the table. First MC check with 100k events.

A good global response diagnostic does not guarantee a reliable tune optimum.

CMS Generator Meeting, July 2026 (internal), slides 3 and 4. Early $N = 200$ diagnostic, separate from the five-seed sampling study.

GP-guided sampling with a common Professor2 fit



Illustrative 2D example; the study has six parameters.

Gaussian Process (GP) guided design

Start from 50 *maximin* points, space-filling and maximising the smallest distance between points. Then add batches of 25, ranking by

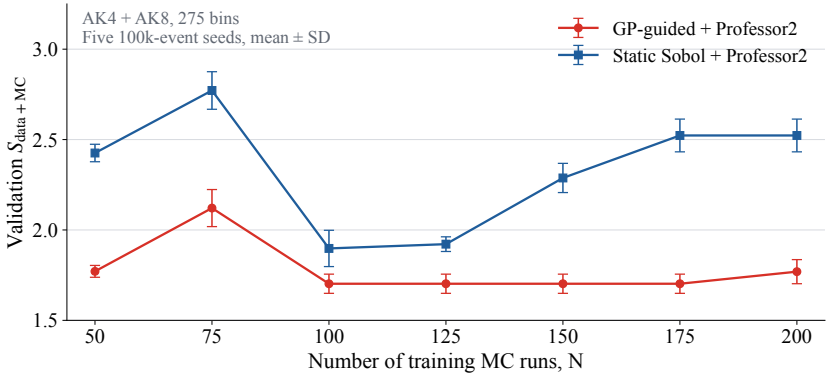
- **exploitation** (low predicted score)
- **uncertainty** (poorly constrained GP)
- **coverage** (far from existing runs)

Static Sobol baseline. Nested coordinates fixed before observing MC responses.
Common fit. Quadratic Professor2 with the same active bins, objective and 8,192 candidates.

Scope. Adaptive GP selection versus Sobol points fixed in advance. Other adaptive methods were not tested.

CMS Generator Meeting, July 2026 (internal). Sampling-method comparison. Ilten et al., JINST 12 (2017) P04028.

Generator validation across the sampling budget

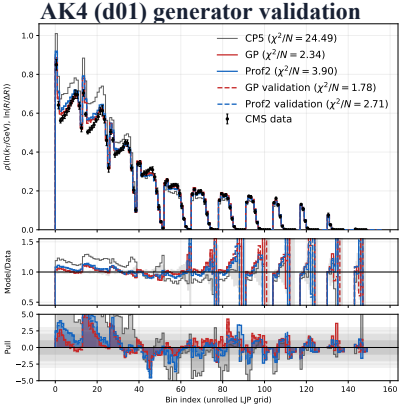


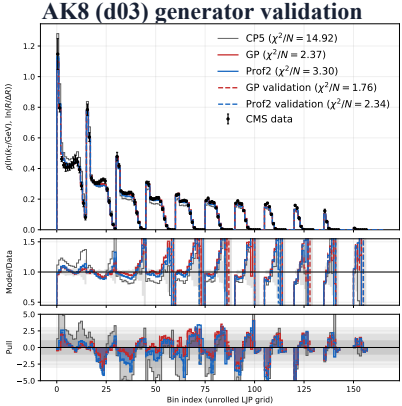
Lower validation scores for GP-guided sampling at every tested budget.

Error bars are the spread of five generator runs at the *same* selected point, so MC statistical noise only and not the spread of repeating the sampling campaign.

July campaign CSV, replotted with a common data+MC score. Mean ± sample SD, five 100k-event seeds.

AK4 and AK8 spectra at $N = 200$





Inspect the distributions, ratios and pulls as well as the global score.

Original July plots. In the legend, “GP” means GP-guided + Professor2 and “Prof2” means static Sobol + Professor2. Score conventions follow the original figures. Numerical comparison in backup B3.

Constrained tuning with Professor2

LJP objective and constraints

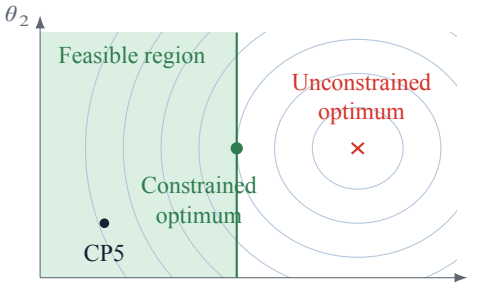
$$\min_{\boldsymbol{\theta}\in\Theta} Q_{\mathrm{LJP}}(\boldsymbol{\theta})$$

$$Q_h(\boldsymbol{\theta}) \le (1+\alpha) Q_h(\mathrm{CP5})$$

for every selected constraint histogram h.

$\alpha = 0$ means no score deterioration relative to CP5, and $\alpha > 0$ an explicit allowed tolerance.

Implementation. Professor2 responses fed to a separate constrained optimiser, SLSQP (sequential least-squares quadratic programming).



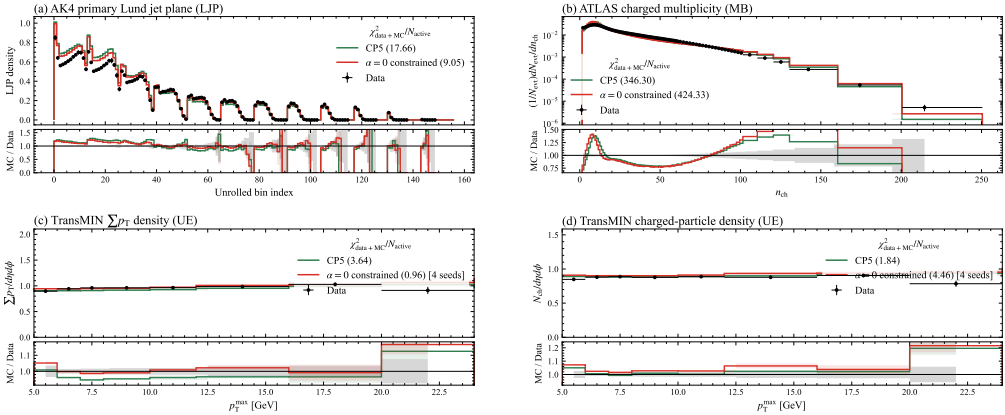
Schematic only, with objective contours and one illustrative constraint boundary

Improve the LJP score while limiting deterioration in selected MB/UE observables.

Constraints act on predicted scores. Independent MC must check them at the selected tune.

Eleven-parameter study with nine selected MB/UE constraint histograms.

Selected observables from the LJP and MB/UE check



Read the MC/Data panels. Panels (a) and (c) move towards unity, (b) and (d) move away.

Data, CP5 and the $\alpha = 0$ candidate. Five seeds (four for constrained UE). 100k events per seed for LJP, 2M for MB/UE.

Scope of the MCNNTUNES comparison

What it is, and why we look at it

MCNNTUNES replaces the polynomial response with a neural network, an alternative route from generator runs to a tune. CMS would like to test it properly. The March forum introduced its two modes, per-bin and inverse (backup B1).

It was already tried in the CMS intrinsic-k_T tuning (GEN-22-001), in a study documented only internally. At that time the inverse model did not yet work well. We are equally open to testing other recent tuning tools.

We test the response model and the selected tune separately.

Main question. If the model predicts unseen MC well, does the selected tune also agree with data in new generator runs?

March forum overview, CMS Tuning Forum, March 2026. Lazzarin, Alioli, Carrazza, arXiv 2010.02213. Detailed model modes retained in backup B1.

How we will compare MCNNTUNES

Check	Use
Inputs	The same generator runs, parameter ranges and active bins
Held-out prediction	MC points not used in training
Pseudo-data closure	Recover a known generating point
Final tune	Generate new events at the selected parameters
Professor baseline	The same inputs and score definition

Based on the public MCNNTUNES workflow.

MCNNTUNES checks for the CMS study

Four checks

1. How well does the model predict unseen MC points?
2. Is the selected tune confirmed with new generator runs?
3. Is the result stable across training seeds and data splits?
4. What does each method cost at the same generator budget?

The public package's pseudo-data benchmark is described in MCNNTUNES advanced usage.

What must stay fixed?

Step	What to check
Sampling	Compare point-selection methods with the same final response model.
Response model	Check errors near the chosen optimum.
Score	List the bins, weights, correlations and MC statistical treatment.
Generator check	Run new events at the selected point and inspect the spectra.
Interpretation	Keep fit quality, model discrepancy and statistical uncertainty separate.

Summary of the LJP study and the checks used here.

Next steps

1. **Collect Professor examples**
   Ask users for reproducible issues, useful diagnostics and successful workarounds.

2. **Repeat the sampling and constraint studies**
   Use new candidate sets and a denser response near the CP5-to-compromise region.

3. **Add the MCNNTUNES results**
   Fix the inputs and score, then validate the selected tune with new generator events.

Not covered today, but active. The DPS tuning analysis is in internal review, and a new effort towards general tunes to eventually supersede the CPX and CHX series is starting.

LJP follow-up at the CMS Generator Meeting, July 2026 (internal).

Conclusions

1. **Professor remains useful in CMS tuning.**
   Sampling, response diagnostics and validation determine how reliably it performs.

2. **Sampling and constraints solve different problems.**
   GP-guided sampling improved point selection. The MB/UE study shows why predicted constraints must be checked with new events.

3. **Apply the same checks to MCNNTUNES.**
   Test held-out prediction, then validate the selected tune with new generator runs.

Discussion

MCNNTUNES model modes (backup)

Per-bin model

$$\boldsymbol{\theta} \;\longmapsto\; \widehat{H}_b(\boldsymbol{\theta})$$

Neural networks approximate the generator response. A minimiser then fits the predicted observables to data.

Inverse model

$$\mathbf{H} \;\longmapsto\; \widehat{\boldsymbol{\theta}}(\mathbf{H})$$

A neural network learns the mapping from observables to parameters. Parameter degeneracies matter when interpreting this inference.

The package provides hyperparameter scans and a benchmark that tunes held-out MC runs treated as pseudo-data.

CMS question. Which approach is useful for the target tuning problem, and what validation supports that choice?

Lazzarin, Alioli, Carrazza, arXiv 2010.02213. MCNNTUNES documentation.

Three different validation questions (backup)

Test	Procedure	What it probes
Response prediction	Predict a withheld MC histogram at a known parameter point	Accuracy of the response model
Pseudo-data closure	Tune to withheld MC and compare with its generating parameters	Ability to recover known parameters
Selected-tune validation	Generate new events at the selected tune and compare with data	Whether the selected point works in generator runs

The LJP five-seed check uses the third procedure. The documented MCNNTUNES benchmark uses the second.

CMS Generator Meeting, July 2026 (internal). MCNNTUNES benchmark documentation.

Sampling comparison at $N = 200$ (backup)

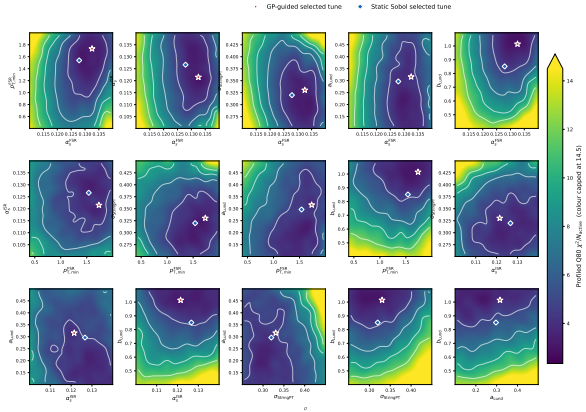
	Professor2 prediction		Five-seed MC validation	
Observable	**GP-guided**	**static Sobol**	**GP-guided**	**static Sobol**
AK4	2.336	3.895	**1.779**	2.707
AK8	2.367	3.302	**1.760**	2.340
Combined	2.352	3.598	**1.769**	2.523

Prediction is the data-only score of the median Professor2 spectrum. Validation is the mean of five per-seed data+MC scores.

GP-guided sampling improves the common Professor2 fit.
The validation ordering holds at tested budgets from 50 to 200.

Independent campaigns must assess sampling variability.
CMS Generator Meeting, July 2026 (internal), slides 17 and 18. Six-parameter method study.

Pairwise projections of the selection objective (backup)



Elongated valleys show where two parameters trade off against each other.

Robust $Q_{0.8}$ objective over the 275 active bins, minimised over the other four parameters in each cell. Active $N = 200$ design, both selected tunes marked. Projections of the fitted response, not confidence regions.