Variational Pseudo Marginal Methods for Jet Reconstruction in Particle Physics

TMLR 2024 JTC Award @ ICLR 2026

Hanming Yang[1*] Antonio Khalil Moretti[1,2*] Sebastian Macaluso[3]
Philippe Chlenski[1] Christian A. Naesseth[4] Itsik Pe'er[1]

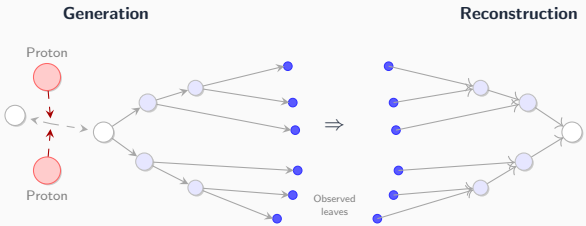
ICLR 2026

[1]Columbia University [2]Spelman College [3]Telefonica Research [4]University of Amsterdam
*Equal contribution

Jet Reconstruction at the LHC

What is a jet?

At CERN's Large Hadron Collider, proton collisions produce **collimated sprays of particles** called jets, the experimental signature of quarks and gluons described by QCD.



Why does it matter?

Reconstructing jets is central to:

- Testing the **Standard Model** of particle physics
- Discovering new particles (Higgs, W/Z bosons)
- Tuning high-fidelity simulators (Pythia, Herwig)

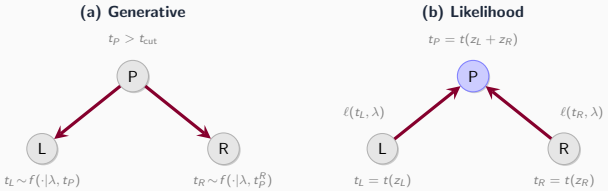
The challenge:

A jet is a **binary tree**. The splitting history τ is **latent**. Only the leaves $\mathbf{X}$ are observed. The number of topologies grows as **(2N-3)!!** which is intractable for $N \geq 15$ particles.

The Ginkgo Generative Model

Why Ginkgo?

Full QCD simulators (Pythia, Herwig) have **intractable likelihoods**. Ginkgo is a semi-realistic model with a **tractable joint likelihood**, enabling exact evaluation and principled probabilistic inference.



Generative process:

A parent node with mass^2 $t_P > t_{\text{cut}}$ recursively splits. Children masses sampled from a truncated exponential:

$$f(t \mid \lambda, t_P) = \frac{\lambda/t_P}{1 - e^{-\lambda}} e^{-\lambda t/t_P}$$

Splitting likelihood at inference:

$$\mathcal{F}(t_L, t_R, \lambda) = \tfrac{1}{4\pi}(1 - F_s)\,\ell(t_L)\,\ell(t_R)$$

Two jet types: **QCD jets** (single λ) and **Heavy Resonance** W boson jets (λ_1, λ_2).

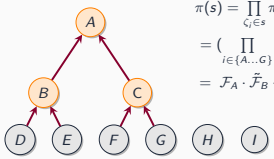
Adapting CSMC and Vcsmc to Jet Reconstruction

Key contribution: recursive likelihood

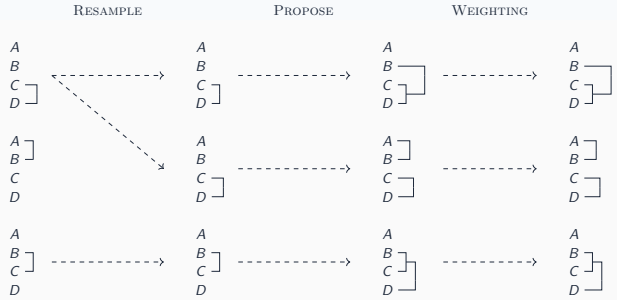
Standard CSMC only sees the last split. We reformulate to capture the **full sub-tree history**:

$$\widetilde{\mathcal{F}}(t_L, t_R, \lambda) = \mathcal{F}(t_L, t_R, \lambda) \times \widetilde{\mathcal{F}}_{\text{left}} \times \widetilde{\mathcal{F}}_{\text{right}}$$

$\widetilde{\mathcal{F}} = 1$ at leaves. Computed via dynamic programming.



$$\pi(s) = \prod_{\zeta_i \in s} \pi(\zeta_i)$$
$$= \Big(\prod_{i \in \{A \ldots G\}} \mathcal{F}_i\Big) \cdot \mathcal{F}_H \cdot \mathcal{F}_I$$
$$= \mathcal{F}_A \cdot \tilde{\mathcal{F}}_B \cdot \tilde{\mathcal{F}}_C = \tilde{\mathcal{F}}_A$$



CSMC builds K partial states over $N-1$ steps:

- **Resample**: keep high-weight states
- **Propose**: merge two trees
- **Reweight**: $w_r^k = \dfrac{\pi(s_r^k)}{\pi(s_{r-1}^{a_{r-1}^k})} \cdot \dfrac{\nu^-}{q}$

Unbiased, consistent estimator: $\widehat{\mathcal{Z}}_{\text{CSMC}} = \prod_{r=1}^{R} \frac{1}{K} \sum_{k=1}^{K} w_r^k \xrightarrow{K \to \infty} \|\pi\|$

Variational Inference & Pseudo-Marginal Framework

Vcsmc: learning topology distributions

Use $\widehat{\mathcal{Z}}_{\text{CSMC}}$ to form an ELBO and jointly optimize the topology posterior Q and parameter λ:

$$\mathcal{L}_{\text{CSMC}} = \mathbb{E}_Q\left[\log \widehat{\mathcal{Z}}_{\text{CSMC}}\right] \leq \log P(\mathbf{X}|\lambda)$$

Vncsmc: nested look-ahead

Enumerates **all** $\binom{N-r}{2}$ **one-step topologies**, subsamples M branch lengths, selects proportional to sub-weights. This is an **exact approximation to the locally optimal proposal**.

Variational Pseudo-Marginal

Place prior $\lambda \sim \log\mathcal{N}(\mu, \Sigma)$ and marginalize jointly over τ *and* λ:

$$P(\tau, \lambda|\mathbf{X}) \propto P(\mathbf{X}|\tau, \lambda)\, P(\tau|\lambda)\, P(\lambda)$$

SMC partial states become **auxiliary variables**, marginalizing recovers correct posterior. **First fully Bayesian treatment** of jets.

Summary of contributions

1. Recursive $\widetilde{\mathcal{F}}$ captures full sub-tree history
2. CSMC for jets: unbiased $\widehat{\mathcal{Z}}$ estimator
3. VCSMC/VNCSMC: VI for τ posterior + λ
4. Variational pseudo-marginal: fully Bayesian $P(\tau, \lambda|\mathbf{X})$

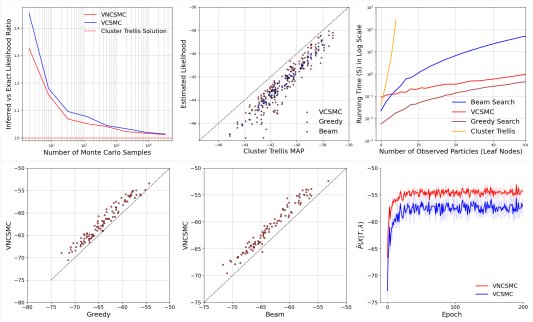
Method	Complexity	Scalable?
CSMC	$\mathcal{O}(KNM)$	**Yes**
NCSMC	$\mathcal{O}(KN^3M)$	Limited
Beam Search	$\mathcal{O}(bN^3 \log N)$	Limited

Results

MAP estimate quality

VNCSMC (K=256) vs. 100 simulated jets:

- **100/100 jets** outperform Greedy Search
- **99/100 jets** outperform Beam Search
- Simultaneously learns λ — baselines need fixed λ
- Returns a **distribution** over topologies, not a single tree



Speed

VCSMC $>10\times$ faster than Beam Search at N=20, scales to N=64+ where Cluster Trellis is intractable.

Convergence

VNCSMC ($K\geq$8) exceeds simulator's own reference likelihood. Even VNCSMC (K=8) outperforms VCSMC (K=256).

Conclusion

We introduce the first adaptation of SMC to jet reconstruction:

1. Recursive splitting likelihood encoding the full sub-tree history
2. CSMC estimator: unbiased, consistent, scalable to $N{=}64+$
3. VCSMC / VNCSMC: variational inference for topology and λ jointly
4. **First fully Bayesian treatment** of all latent variables in jet reconstruction
5. New connections between VSMC and pseudo-marginal MCMC

Poster & paper