
Multimodal Generative Flows for LHC Jets

Darius A. Faroughy
NHETC, Rutgers University

Manfred Opper
TU Berlin

César Ojeda
University of Potsdam

Abstract

Generative modeling of high-energy collisions at the Large Hadron Collider (LHC) offers a data-driven route to simulations, anomaly detection, among other applications. A central challenge lies in the hybrid nature of *particle-cloud* data: each particle carries continuous kinematic features and discrete quantum numbers such as charge and flavor. We introduce a transformer-based *multimodal flow*¹ that extends flow-matching with a continuous-time Markov jump bridge to jointly model LHC jets with both modalities. Trained on CMS Open Data, our model can generate high fidelity jets with realistic kinematics, jet substructure and flavor composition.

1 Introduction

The Large Hadron Collider (LHC) at CERN produces billions of proton–proton collisions per second, reconstructed into final-state particles by multi-layered detectors. Among the many objects emerging from hadronic collisions, *jets*—collimated sprays of localized high-energy particles—play a central role in both QCD studies and searches for new physics. As *particle clouds*, jets have become a testbed for modern generative models. For example, they can be used to learn background distributions directly from data for resonant anomaly detection, bypassing the limitations of imperfect Monte Carlo simulations and avoiding reliance on high-level, hand-crafted observables [Buhmann et al. (2024)]. These models can also act as tractable high-dimensional density estimators, offering a principled framework for evaluating how closely jet classifiers approach the theoretical optimum [Geuskens et al. (2024)].

Dynamics-based frameworks such as diffusion [Sohl-Dickstein et al. (2015); Song et al. (2020)] and flow-matching [Albergo & Vanden-Eijnden (2022); Lipman et al. (2022)] have recently set new benchmarks for jet generation and now underpin state-of-the-art foundation models for particle physics [Mikuni & Nachman (2025)]. However, these methods operate solely on continuous spaces. This is inadequate for jets, where each particle carries both continuous kinematics and categorical attributes like charge and flavor. The use of de-quantization methods or modeling these modalities separately risks distorting physically meaningful correlations. Accurate modeling therefore requires a framework that jointly treats both continuous and discrete modalities. Multimodal flows have recently gained attention in other scientific fields [Campbell et al. (2024)], see related work in App A.

We propose a multimodal generative framework that integrates continuous flow-matching with a continuous-time Markov jump process for the discrete dynamics. This yields a unified probabilistic path over hybrid spaces capable of generating both kinematics and quantum numbers for jet constituents. Trained on CMS Open Data, our model accurately reproduces the kinematics, substructure and flavor content of real world jets. We show that our model, when equipped with a multimodal particle transformer architecture, with mode-specific and fused encoders can produce state of the art results on the ASPENOPENJETS dataset introduced by Amram et al. (2024).

¹code repository github.com/dfarroughy/Multimodal-flows

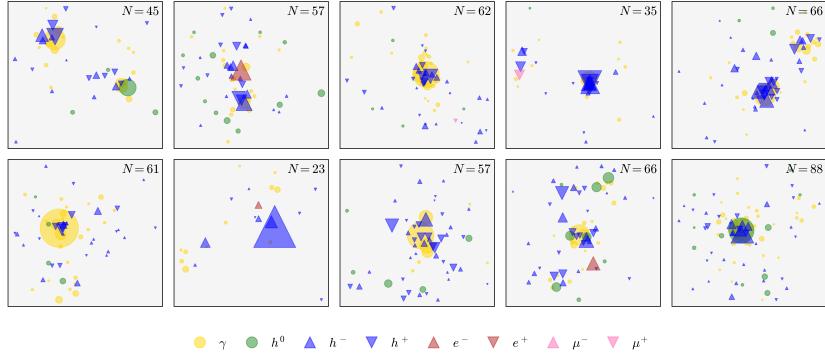


Figure 1: Visualizations of 10 CMS jets in the (η, ϕ) -plane taken from the ASPENOPENJETS dataset. Each particle cloud is centered around the jet axis. The size of each marker is proportional to the particle’s transverse momentum (p_T), while the shape and color encode electric charge and flavor, respectively. The upper right corner of each panel indicates the total number of constituents in the respective jet.

2 CMS Open Data

In this work we are interested in training generative models on real LHC jets. We use the recently released ASPENOPENJETS (AOJ) dataset derived from 13 TeV proton-proton collisions recorded by CMS in 2016. Each jet is represented as a particle-cloud $\mathbf{z} = \{z^d\}_{d=1}^D$ with up to $D = 150$ constituents, where each particle is described by (continuous) kinematic features in hadronic coordinates and a (categorical) *flavor* token: $z^d \equiv (x^d, k^d) \in \mathbb{R}^3 \otimes \mathcal{F}$ with $x^d \in (p_T, \Delta\eta, \Delta\phi)$ and $k^d \in \mathcal{F} = \{\gamma, h^0, h^-, h^+, e^-, e^+, \mu^-, \mu^+\}$. Here $p_T = \sqrt{p_x^2 + p_y^2}$ is the transverse momentum, and $\Delta\eta, \Delta\phi$ are coordinates relative to the jet axis. The possible flavors are: photons (γ), electrically charged hadrons (h^- , h^+), electrically neutral hadrons (h^0), electrons (e^-), positrons (e^+), muons (μ^-) and anti-muons (μ^+), corresponding to particle species and charges reconstructed by the CMS tracking system and calorimeters. Representative particle clouds from the AOJ dataset are shown in Fig. 1. The flavor composition is strongly imbalanced: photons and charged hadrons make up $\sim 90\%$ of all constituents (in almost equal proportions), neutral hadrons $\sim 10\%$, while leptons occur only at the per-mille level.

3 Multimodal flows for particle clouds

We now describe our extension of flow-matching over the hybrid space $\mathbb{R}^3 \otimes \mathcal{F}$. We provide extensive supplementary material to this section in App. B. We denote by $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{k}_t)$, the time-dependent path of the jet that transforms an arbitrary source data $\mathbf{z}_0 = (\mathbf{x}_0, \mathbf{k}_0) \sim \mu$ at $t = 0$ into the target data $\mathbf{z}_1 = (\mathbf{x}_1, \mathbf{k}_1) \sim \nu$ at $t = 1$ and $P_t(\mathbf{z})$ the corresponding probability path satisfying

$$\partial_t P_t(\mathbf{z}_t) = -\nabla_{\mathbf{x}} \cdot [\mathbf{u}_t(\mathbf{z}_t) P_t(\mathbf{z}_t)] + \sum_{j \neq k_t} [\mathbf{W}_t(\mathbf{z}_t, j) P_t(j) - \mathbf{W}_t(j, \mathbf{z}_t) P_t(\mathbf{z}_t)], \quad (1)$$

and subject this the boundary conditions $P_0 = \mu(\mathbf{x}_0, \mathbf{k}_0)$ and $P_1 = \nu(\mathbf{x}_1, \mathbf{k}_1)$. The first term is the familiar continuity equation for continuous densities where \mathbf{u}_t is the velocity field acting on the particle kinematics $\mathbf{x} \in \mathbb{R}^3$, while the remaining terms describe a continuous time Markov jump bridge generating discrete transitions between flavor tokens $\mathbf{k} \in \mathcal{F}$ with a jump rate matrix \mathbf{W}_t . Following the conditional flow-matching framework, we introduce conditional quantities and write the path as a marginalization over the boundary data:

$$P_t(\mathbf{z}_t) = \sum_{\mathbf{k}_0, \mathbf{k}_1} \int d\mathbf{x}_0 d\mathbf{x}_1 \mu(\mathbf{x}_0, \mathbf{k}_0) \nu(\mathbf{x}_1, \mathbf{k}_1) \prod_{d=1}^D p_t(x_t^d | x_0^d, x_1^d) q_t(k_t^d | k_0^d, k_1^d). \quad (2)$$

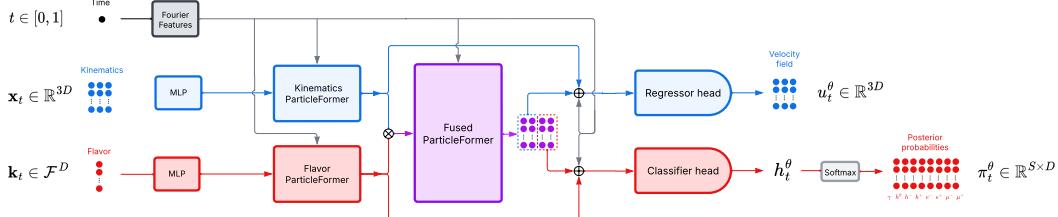


Figure 2: Multi-modal particle transformer architecture. Additional details are provided App. C.

Here, to ensure tractability we impose a complete factorization over particles and both modalities, where the continuous probability density $p_t(x^d|x_0^d, x_1^d)$ and discrete probability mass function $q_t(k^d|k_0^d, k_1^d)$, satisfy separate conditional dynamics. Application of this marginalization trick yields a velocity field u_t and a jump rate W_t expressed in terms of the expectations of the conditional velocities and rates with respect to the *posterior probability* distributions:

$$\mathbf{u}_t(\mathbf{z}_t) = \mathbb{E}_{\pi_t(\mathbf{z}_0, \mathbf{z}_1 | \mathbf{z}_t)} \mathbf{u}_t(\mathbf{x}_t | \mathbf{x}_0, \mathbf{x}_1) \quad (3)$$

$$W_t(j, \mathbf{z}_t) = \mathbb{E}_{\pi_t(\mathbf{z}_0, \mathbf{z}_1 | \mathbf{z}_t)} W_t(k_t, j | k_0, k_1), \quad (4)$$

where the posterior follows $\pi_t(\mathbf{z}_0, \mathbf{z}_1 | \mathbf{z}) = p_t(\mathbf{x} | \mathbf{x}_0, \mathbf{x}_1) q_t(\mathbf{k} | \mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{z}_0) \nu(\mathbf{z}_1) / P_t(\mathbf{z})$ from Bayes theorem. Although each component evolves independently under the conditional process, non-trivial correlations between particles and their respective modalities emerge for the generative process (2).

Conditional dynamics The next step is to specify the conditional dynamics over $\mathbb{R}^3 \otimes \mathcal{F}$. For the continuous modality, the standard choice is to take a *uniform flow* that transports source data into target data through straight paths with constant velocities [Liu et al. (2022)]. For the discrete modality we propose a multivariate generalization of the *random telegraph process* [Gardiner (2010)], a continuous-time Markov process originally used to model noise in binary communication channels. As shown in App. B.2, the resulting velocity field and jump rate matrix are given by:

$$u_t^d(x_t^d | x_0^d, x_1^d) = x_1^d - x_0^d \quad (5)$$

$$W_t^d(k_t^d = i, j^d = j | k_1^d = k) = 1 + \frac{S \omega_t}{1 - \omega_t} \delta_{ik} + \omega_t \delta_{jk}. \quad (6)$$

where S is the token vocabulary size, $\omega_t \equiv \exp(-S\beta(1-t))$ and β is a stochasticity hyperparameter that controls the frequency of jumps per unit time.

Multimodal objective Substituting these solutions into Eqs. (3) and (4), the expectation of the velocity field cannot be expressed in closed form and must be approximated from data with the conditional flow-matching objective by regressing the conditional vector field with a parametric function $u_t^\theta(\mathbf{z}_t)$. By contrast, the expectation over the discrete jump rates is fully tractable and yields the explicit expression

$$W_t^d(k^d = i, j^d = j) = 1 + \frac{S \omega_t}{1 - \omega_t} \pi_t(k_1^d = i | j^d = j) + \omega_t \pi_t(k_1^d = j | j^d = j), \quad (7)$$

where π_t denotes the posterior distribution for particle d over the final state token i conditioned on the intermediate state j at time t . Therefore, if the posterior π_t can be approximated with a parametric estimate π_t^θ from the data, one can directly compute the rates via (7). In this case the posterior learning problem is equivalent to a multi-class classification task. We thus introduce a time-dependent classifier h_t^θ such that the vector of posterior probabilities is given by the *softmax* function $\pi_t^\theta = \text{softmax}(h_t^\theta)$ and train the classifier function by minimizing the cross-entropy loss.

We train using the flow-matching mean-square error (MSE) loss for the particle kinematics and the cross-entropy (CE) loss for the particle flavor tokens. Following the approach of Kendall et al. (2018), these two objectives are combined into a single weighted loss:

$$\mathcal{L}_{\text{MMF}} = \mathbb{E}_{t, (\mathbf{z}_0, \mathbf{z}_1), \mathbf{z}_t} \left[\frac{\|u_t^\theta(\mathbf{z}_t) - u_t(\mathbf{x}_t | \mathbf{x}_0, \mathbf{x}_1)\|^2}{2(\sigma_t^1)^2} - \frac{\log h_t^\theta(\mathbf{z}_t, \mathbf{k}_1)}{2(\sigma_t^1)^2} + \log(\sigma_t^1 \sigma_t^2) \right]. \quad (8)$$

In the expectation, time is drawn uniformly $t \sim \mathcal{U}[0, 1]$ over the unit interval, pairs of source and target points drawn from the coupling $(\mathbf{z}_0, \mathbf{z}_1) \sim \mu \otimes \nu$ and $\mathbf{z}_t \sim P_t(\cdot | \mathbf{z}_0, \mathbf{z}_1)$. In contrast to Kendall et al. (2018), where the uncertainty weights σ_i are fixed trainable scalars, we promote them to time-dependent functions σ_t^i . In practice, we parametrize the weights via $\sigma_t^i = \exp(-w_t^i)$, where w_t is the output of an *uncertainty network* that we discard during inference. This allows the relative weighting between modalities to adapt dynamically along the generative path, enabling the model to prioritize different objectives at different stages of the evolution. Besides improving the convergence of the training, this formulation eliminates the need for manually tuning loss weights.

Multimodal ParticleFormer To learn the continuous and discrete modalities together, we approximate the conditional velocity field and the posterior classifier function with a single permutation equivariant neural network $u_t^\theta \otimes h_t^\theta$. The overall architecture, depicted in Fig. 2, has three components: 1) two mode-specific encoders, 2) a *fused* encoder, and 3) two task-specific heads. All three encoders consist of non-causal particle transformers [Qu et al. (2022)] with stacked multi-head self-attention blocks. The *regressor head* predicts the continuous-valued vector field u_t^θ for the MSE loss, while the *classifier head* outputs the logits h_t^θ for the CE loss. For more details on the architecture see App. C.

Sampling Once the model is trained, new samples can be generated from a source input \mathbf{z}_0 by simulating the associated dynamics to the probability flow equation (1). This entails numerically solving the joint dynamics for both the continuous and discrete components. For the continuous dynamics we solve the ordinary differential equation $\dot{\mathbf{x}}_t = \mathbf{u}_t^\theta(\mathbf{x}_t)$ with Euler’s method using discrete time-steps of size Δt . The discrete dynamics of the flavor tokens follows the multivariate telegraph process governed by the rate \mathbf{W}_t^θ derived from the trained posterior $\pi_t^\theta = \text{softmax}(h_t^\theta/T)$. Here we have introduced a temperature scaling hyperparameter T [Guo et al. (2017)] that helps improving sampling quality. To efficiently simulate this stochastic process, we employ τ -leaping [Gillespie (2001); Campbell et al. (2022)], a well-established approximation method widely used in chemical reaction kinetics. More details on the sampling algorithm can be found in App. D.

4 Experiments

Aspen Open Jets We demonstrate that our transformer-based multimodal flow (MMF) is capable of generating the particle kinematics and flavor composition of real-world jets from CMS open data. Since our work consists of a proof-of-concept, we do not attempt a full optimization of the architectures and only train our models on a subset of the `AspenOpenJets` dataset. Our training dataset consists of 1.25 million AOJ jets, with 1 million jets for training and 250K jets for validation. For the source data $\mathbf{z}_0 = \{\mathbf{z}_0^d\}_{d=1}^D$, we draw point-cloud noise from the distribution $\mathbf{z}_0^d = (x_0^d, k_0^d) \sim \nu \equiv \mathcal{N}(0, 1) \otimes \mathcal{U}(p)$ where \mathcal{N} denotes the standard Gaussian over \mathbb{R}^3 and \mathcal{U} represents the uniform categorical distribution over the token space \mathcal{F} with probability parameter $p = 1/S$. We compare our proposed model to EPiC-FM, a permutation equivariant flow-matching model with deep-sets architecture designed for particle cloud data that achieves state-of-the-art results on simulated jets [Buhmann et al. (2023b,a); Birk et al. (2023)]. Our training setup is described in App. E. We generate samples with 270K jets using a temperature scaling of $T = 0.85$ and solve the dynamics with $\Delta t = 0.001$ time-step size. To quantify the performance of the continuous component of our generator we compute the jet-level kinematics (p_T, m, η, ϕ) and the N -subjettiness ratios τ_{21}, τ_{32} [Thaler & Van Tilburg (2011)] as probes for jet substructure. For the discrete component we examine the particle flavor multiplicities $N^k, k \in \mathcal{F}$. Finally, to test the model capacity in reproducing cross-modal correlations we compute the *jet charge* $\mathcal{Q} = \sum_{i \in \text{jet}} Q_i p_{T,i}/p_T^{\text{jet}}$, a non-trivial aggregate of the particles electric charge Q_i and transverse momentum $p_{T,i}$. These distributions, and the corresponding Wasserstein-1 distances to the AOJ truth, are provided in Fig. 3 and Table 1, respectively.

Discussion The results show that our MMF model outperforms the EPiC-FM baseline in several respects: (i) jets are more accurately centered in the η - ϕ plane, (ii) jet substructure observables are reproduced with higher fidelity, and (iii) jet charge distributions agree more closely with data,

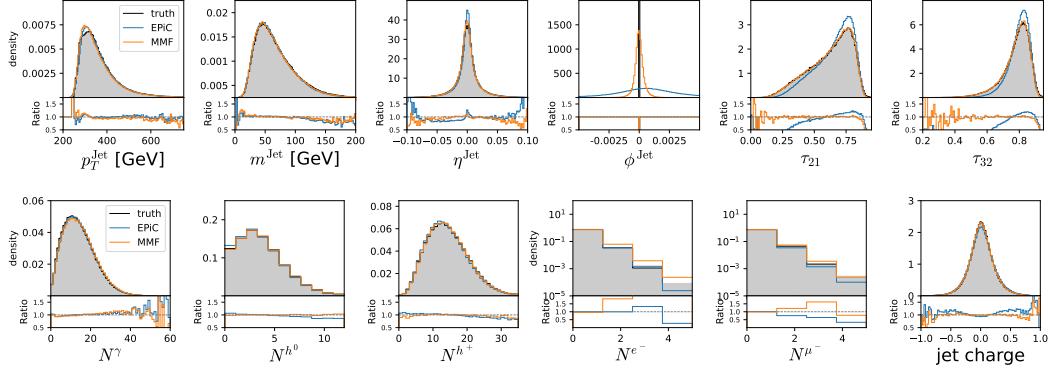


Figure 3: Performance comparison between generated samples from our particle transformer MMF model (orange) and the EPiC-FM baseline (blue) for various high-level jet observables. The corresponding Wasserstein distance between the generated and test distributions are shown in Table 1.

Table 1: The Wasserstein distances $W_1^{\mathcal{O}}$ computed between the generated data and the test data for each jet observables \mathcal{O} . Lower values are better.

Model	$W_1^{p_T}$	W_1^m	W_1^η	W_1^ϕ	$W_1^{\tau_{21}}$	$W_1^{\tau_{32}}$	W_1^Q
EPiC-FM	0.92	1.63	1.2×10^{-3}	2.8×10^{-3}	3.1×10^{-2}	1.8×10^{-2}	9.5×10^{-3}
MMF (ours)	4.64	1.26	6.3×10^{-4}	2.3×10^{-4}	2.3×10^{-3}	2.8×10^{-3}	1.4×10^{-3}

Model	$W_1^{N^\gamma}$	$W_1^{N^{h^0}}$	$W_1^{N^{h^-}}$	$W_1^{N^{h^+}}$	$W_1^{N^{e^-}}$	$W_1^{N^{e^+}}$	$W_1^{N^{\mu^-}}$	$W_1^{N^{\mu^+}}$
EPiC-FM	0.23	0.10	0.28	0.23	5.6×10^{-4}	6.8×10^{-4}	2.6×10^{-3}	3.2×10^{-3}
MMF (ours)	0.34	0.01	0.09	0.10	5.7×10^{-2}	5.6×10^{-2}	4.3×10^{-2}	4.3×10^{-2}

indicating superior modeling of cross-modal correlations. We attribute these improvements to the particle transformer architecture, which better captures inter-particle correlations than the Deep Sets-based EPiC encoder². On the other hand, EPiC-FM provides good modeling of all flavor multiplicities, successfully reproducing dominant (γ , h^\pm), subdominant (h^0), and even rare (e^\pm , μ^\pm) classes. MMF outperforms EPiC-FM for the dominant and subdominant flavors but slightly underperforms for the rare leptons, whose per-mille frequency in the training data remains a limiting factor. We hypothesize that residual stochastic noise from the τ -leaping sampler affects these minority classes near the time endpoint $t \approx 1$. If true, this limitation could be mitigated by applying a post-sampling calibration to the generated jets. Finally, we note that the choice of temperature scaling is critical: while $T = 0.85$ yields unbiased multiplicities, when scanning over T during our experiments, we found that departures from this value systematically distort the neutral-hadron distribution N^{h^0} . This highlights the importance of this hyperparameter in our setup.

Acknowledgments and Disclosure of Funding

Darius Faroughy was supported by DOE grant DOE-SC0010008. César Ojeda was supported by the Deutsche Forschungsgemeinschaft (DFG)– Project-ID 318763901– SFB1294. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award HEP-ERCAP0027491

²Interestingly, both methods have some trouble capturing the irregular shape of the p_T peak, with the MMF model performing slightly worst than EPiC-FM.

References

- Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.
- Amram, O., Anzalone, L., Birk, J., Faroughy, D. A., Hallin, A., Kasieczka, G., Krämer, M., Pang, I., Reyes-Gonzalez, H., and Shih, D. Aspen open jets: Unlocking lhc data for foundation models in particle physics. *arXiv preprint arXiv:2412.10504*, 2024.
- Araz, J. Y., Mikuni, V., Ringer, F., Sato, N., Acosta, F. T., and Whitehill, R. Point cloud-based diffusion models for the electron-ion collider. *arXiv preprint arXiv:2410.22421*, 2024.
- Birk, J., Buhmann, E., Ewen, C., Kasieczka, G., and Shih, D. Flow matching beyond kinematics: Generating jets with particle-id and trajectory displacement information. *arXiv preprint arXiv:2312.00123*, 2023.
- Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- Buhmann, E., Ewen, C., Faroughy, D. A., Golling, T., Kasieczka, G., Leigh, M., Quétant, G., Raine, J. A., Sengupta, D., and Shih, D. Epic-ly fast particle cloud generation with flow-matching and diffusion. *arXiv preprint arXiv:2310.00049*, 2023a.
- Buhmann, E., Kasieczka, G., and Thaler, J. EPiC-GAN: Equivariant point cloud generation for particle jets. *SciPost Phys.*, 15(4):130, 2023b. doi: 10.21468/SciPostPhys.15.4.130.
- Buhmann, E., Ewen, C., Kasieczka, G., Mikuni, V., Nachman, B., and Shih, D. Full phase space resonant anomaly detection. *Phys. Rev. D*, 109(5):055015, 2024. doi: 10.1103/PhysRevD.109.055015.
- Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 28266–28279. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b5b528767aa35f5b1a60fe0aaeca0563-Paper-Conference.pdf.
- Campbell, A., Yim, J., Barzilay, R., Rainforth, T., and Jaakkola, T. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.
- Chopin, N., Fulop, A., Heng, J., and Thiery, A. H. Computational doob h-transforms for online filtering of discretely observed diffusions. In *International Conference on Machine Learning*, pp. 5904–5923. PMLR, 2023.
- Fitzsimmons, P., Pitman, J., and Yor, M. Markovian bridges: construction, palm interpretation, and splicing. In *Seminar on Stochastic Processes*, 1992, pp. 101–134. Springer, 1992.
- Gardiner, C. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer Berlin Heidelberg, 2010. ISBN 9783642089626. URL <https://books.google.com/books?id=321EuQAACAAJ>.
- Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R. T., Synnaeve, G., Adi, Y., and Lipman, Y. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37:133345–133385, 2024.
- Geuskens, J., Gite, N., Krämer, M., Mikuni, V., Mück, A., Nachman, B., and Reyes-González, H. The fundamental limit of jet tagging. *arXiv preprint arXiv:2411.02628*, 2024.
- Gillespie, D. T. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 115(4):1716–1733, 2001.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

- Holderrieth, P., Havasi, M., Yim, J., Shaul, N., Gat, I., Jaakkola, T., Karrer, B., Chen, R. T., and Lipman, Y. Generator matching: Generative modeling with arbitrary markov processes. *arXiv preprint arXiv:2410.20587*, 2024.
- Johnson, J. M. and Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54, 2019.
- Kendall, A., Gal, Y., and Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Lin, H., Li, S., Ye, H., Yang, Y., Ermon, S., Liang, Y., and Ma, J. Tfg-flow: Training-free guidance in multimodal generative flow. *arXiv preprint arXiv:2501.14216*, 2025.
- Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T., Lopez-Paz, D., Ben-Hamu, H., and Gat, I. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.
- Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL <https://arxiv.org/abs/2209.03003>.
- Mikuni, V. and Nachman, B. Solving key challenges in collider physics with foundation models. *Phys. Rev. D*, 111(5):L051504, 2025. doi: 10.1103/PhysRevD.111.L051504.
- Qu, H., Li, C., and Qian, S. Particle Transformer for Jet Tagging. 2 2022.
- Ren, Y., Chen, H., Rotkoff, G. M., and Ying, L. How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. *arXiv preprint arXiv:2410.03601*, 2024.
- Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.
- Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.
- Thaler, J. and Van Tilburg, K. Identifying Boosted Objects with N-subjettiness. *JHEP*, 03:015, 2011. doi: 10.1007/JHEP03(2011)015.
- Winkler, L., Richter, L., and Opper, M. Bridging discrete and continuous state spaces: Exploring the ehrenfest process in time-continuous diffusion models. *arXiv preprint arXiv:2405.03549*, 2024.

Supplementary Material

A Related Work

Generative models that independently handle discrete and continuous variables have been explored since the inception of research in both denoising and diffusion frameworks, with various methodologies attempting to relate these approaches. Winkler et al. (2024) show that a limit for Ehrenfest processes defined on discrete spaces converges to an Ornstein-Uhlenbeck dynamic. Further extending the theory, Ren et al. (2024) establish a stochastic integral formulation of discrete diffusion models, generalizing the Poisson random measure through a Lévy-type integral. The prevalence of multimodal flows in the natural sciences has spurred additional multimodal solutions akin to our work [Lin et al. (2025)]. Purely discrete flows were developed by Gat et al. (2024), while in protein design, Campbell et al. (2024) introduced a fully multimodal methodology. Their approach differs from ours in selecting conditional paths in an ad hoc manner, directly interpolating within probability space. Conversely, we introduce a more general strategy allowing diverse probabilistic paths by leveraging Markov bridges, as presented by Chopin et al. (2023). Finally, a general extension of flow matching theory to a broader family of stochastic process generators was proposed by Holderrieth et al. (2024); Lipman et al. (2024). Our method can be viewed as a special instance of this broader framework, where the probability path is derived via a Markov bridge applied to a specific type of reference process.

B Generative modeling with bridge processes

In dynamic generative modeling, one has access to samples from a target distribution ν , and the goal is to generate novel unseen samples from that same distribution. To generate new samples, we start with a tractable source distribution μ (e.g., a Gaussian) and transform its samples through a deterministic or stochastic generative process $\{\mathbf{z}_t\}_{t \in [0,1]}$, such that by the final time $t = 1$, the distribution of the transformed variables \mathbf{z}_1 match the desired distribution ν . Any such transformation has an associated probability path $P_t(\mathbf{z})$, and the objective is to construct the transformation such that the probability path ensures $P_0 = \mu$ and $P_1 = \nu$. In the following, we will refer to the probability path that complies to this boundary conditions as the *target probability path*³.

Stochastic transformations can be specified as stochastic processes through an infinitesimal generator \mathcal{L}_t or its adjoint operator and the corresponding Fokker–Planck or Master equations, whereas deterministic transformations can be achieved with a *flow* ψ_t specified in turn with a velocity field u_t . The family of dynamic generative models, constitute a group of methodologies that are able to construct these transformations via neural networks approximations to the desired target generators \mathcal{L}_t^θ or vector fields u_t^θ . In particular, the flow matching family achieves this through a conditional strategy, whereby one designs a conditional probability path that, after marginalization, recovers the target path. Crucially, the conditional path is defined to ensure access to tractable conditional generators or velocities. The key insight of this methodology is that the target generator or velocity field can be constructed or learned by properly averaging the conditional generators. The average however, is performed over a posterior distribution from the conditional paths.

In our formulation, we closely follow the flow matching methodology of Albergo & Vanden-Eijnden (2022); Lipman et al. (2022) and start by constructing probability paths that follow a prescribed continuity equation. We then construct conditional probability paths with the help of *reference process* with known generators that enables the construction of *bridges* between point samples, ensuring transformations from \mathbf{z}_0 to \mathbf{z}_1 . In the context of particle clouds for jets, we will assume that $\mathbf{z} = (\mathbf{x}, \mathbf{k})$ is a collection of continuous (\mathbf{x}) and discrete (\mathbf{k}) vectorial random variables. And in the following we will show the construction for \mathbf{x} and \mathbf{k} separately.

³Contrary to the diffusion literature, where the generative process is the backward process running in reverse-time, we follow throughout this paper the flow-based convention where the generative process is forward-time.

B.1 Continuous random variables

First, we consider the continuous case with $z \equiv x$ and $P_t \equiv p_t$. Our goal is to construct probability paths that satisfy the continuity equation:

$$\partial_t p_t(\mathbf{x}) = -\nabla \cdot [\mathbf{u}_t(\mathbf{x}) p_t(\mathbf{x})] \quad (9)$$

subject to the boundary conditions $p_1 = \nu$ and $p_0 = \mu$. It is known that such continuity equations specify the transformation of the random variables $X_0 \sim \mu$ with a flow $\mathbf{x}_t = \psi_t(\mathbf{x}_0)$ whose dynamics are given by:

$$\frac{d\psi_t(\mathbf{x})}{dt} = \mathbf{u}_t(\psi_t(\mathbf{x})) , \text{ with } \psi_{t=0}(\mathbf{x}) = \mathbf{x}. \quad (10)$$

Where \mathbf{u}_t is the target velocity field. Using the flow ψ_t the marginal PDFs can be obtained via the push back formula $p_t(\mathbf{x}) = [\psi_t \# p](\mathbf{x})$.

Conditional Flow-Matching Since one does not have access to close forms of \mathbf{u}_t that solve equation (9) given the boundary conditions, we will construct it by first introducing conditional probability paths $p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1)$ that are able to obtain the target probability path p_t by marginalizing over the target ν and data distribution μ :

$$p_t(\mathbf{x}) = \int p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) \mu(\mathbf{x}_0) \nu(\mathbf{x}_1) d\mathbf{x}_0 d\mathbf{x}_1 \quad (11)$$

the desired boundary conditions of the target path are obtained if the conditional fulfill:

$$\lim_{t \rightarrow 0} p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) = \delta(\mathbf{x} - \mathbf{x}_0) \quad (12)$$

$$\lim_{t \rightarrow 1} p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) = \delta(\mathbf{x} - \mathbf{x}_1) \quad (13)$$

here $\delta(\cdot)$ denotes the Dirac function. Many such conditional paths can be constructed with these boundary conditions, in the standard flow matching methodology, this is achieved by linearly interpolating between the end points:

$$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0 \quad (14)$$

Hence, for this construction, we obtain *Dirac* probability paths $p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) = \delta(\mathbf{x} - \mathbf{x}_t)$. One can show that such probability paths are generated by the following conditional vector field:

$$\mathbf{u}_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) = \mathbf{x}_1 - \mathbf{x}_0 \quad (15)$$

Now one can prove [Albergo & Vanden-Eijnden (2022); Lipman et al. (2022)] that the desired target vector field, can be obtained from the conditional by:

$$\mathbf{u}_t(\mathbf{x}) = \mathbb{E}_{\rho_t(\mathbf{x}_0, \mathbf{x}_1 | \mathbf{x})} [\mathbf{u}_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1)] \quad (16)$$

where the expectation is performed with respect to the *posterior probability* density of end point pairs $(\mathbf{x}_0, \mathbf{x}_1)$ conditioned on the intermediate point $\mathbf{x}_t = \mathbf{x}$ at time t . This posterior is obtained by applying Bayes formula:

$$\rho_t(\mathbf{x}_0, \mathbf{x}_1 | \mathbf{x}) = \frac{p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) \mu(\mathbf{x}_0) \nu(\mathbf{x}_1)}{p_t(\mathbf{x})}. \quad (17)$$

In order to compute (16) we can use the widely known fact that the conditional expectation can be expressed as the minimizer of an appropriate least-square error problem equivalent to a nonlinear regression problem. We use a neural network approximation $\mathbf{u}_t^\theta(\mathbf{x}_t)$ to the minimizer, where θ corresponds to the parameters of the network. To be precise, the conditional expectation is obtained as the minimum of the mean square error (MSE) loss:

$$\mathcal{L}_{CFM} = \mathbb{E}_{t, (\mathbf{x}_0, \mathbf{x}_1), \mathbf{x}_t} \|\mathbf{u}_t^\theta(\mathbf{x}_t) - \mathbf{u}_t(\mathbf{x}_t | \mathbf{x}_0, \mathbf{x}_1)\|^2 \quad (18)$$

The expectation runs over $t \sim \mathcal{U}(0, 1)$, $(\mathbf{x}_0, \mathbf{x}_1) \sim \mu \otimes \nu$ and $\mathbf{x}_t \sim p_t(\cdot | \mathbf{x}_0, \mathbf{x}_1)$.

B.2 Discrete random variables

We now consider the case where the data $\mathbf{z} \equiv \mathbf{k}$, $P_t \equiv q_t$, takes values in a discrete space $\mathbf{k} \in \{0, 1, \dots, S\}^D$, and where the source is obtained from a simple probability mass function $\mathbf{k}_0 \sim \mu$ and one has, as before, access to the target distribution via data samples from ν . Our goal is to construct a target probability path $q_t(\mathbf{k})$ between $\mu(\mathbf{k})$ and $\nu(\mathbf{k})$. Similar to (9) we start by imposing a continuity equation to the target probability path, which for the discrete variable case corresponds to the *Master Equation*:

$$\partial_t q_t(\mathbf{k}) = \sum_{j \neq \mathbf{k}} [\mathbf{W}_t(\mathbf{k}, j) q_t(j) - \mathbf{W}_t(j, \mathbf{k}) q_t(\mathbf{k})] \quad (19)$$

and subject this equation to the boundary conditions $q_0(\mathbf{k}) = \mu$ and $q_1(\mathbf{k}) = \nu$. This is a type of stochastic process in which transitions, or *jumps*, between discrete states occur at continuous random times. The process is fully defined by rate matrices $\mathbf{W}_t \in \mathbb{R}^{(S+1)^D \times (S+1)^D}$, where for $\mathbf{k} \neq j$, $\mathbf{W}_t(\mathbf{k}, j) dt$ equals the average number of jumps from state j to state \mathbf{k} at time t occurring during an infinitesimal time window dt . Formally:

$$\mathbf{W}_t(\mathbf{k}, j) \doteq \lim_{\Delta t \rightarrow 0} \frac{q_{t+\Delta t|t}(\mathbf{k}|j) - \delta_{k,j}}{\Delta t} \quad (20)$$

where we have defined the transition probability by $q_{t|s}(\cdot|\cdot)$ and the symbol q denotes the computation of probabilities with respect to q . Since transition probabilities are normalized, one has $\mathbf{W}_t(\mathbf{k}, \mathbf{k}) = -\sum_{\mathbf{k}' \neq \mathbf{k}} \mathbf{W}_t(\mathbf{k}', \mathbf{k})$.

Now, since the desired target rate \mathbf{W}_t that fulfills (19) subject to the boundary conditions is unknown, we proceed as before by introducing a conditional probability path such that:

$$q_t(\mathbf{k}) = \sum_{\mathbf{k}_0, \mathbf{k}_1} q_t(\mathbf{k}|\mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{k}_0) \nu(\mathbf{k}_1), \quad (21)$$

$$\lim_{t \rightarrow 0} q_t(\mathbf{k}|\mathbf{k}_0, \mathbf{k}_1) = \delta_{\mathbf{k}, \mathbf{k}_0}, \quad (22)$$

$$\lim_{t \rightarrow 1} q_t(\mathbf{k}|\mathbf{k}_0, \mathbf{k}_1) = \delta_{\mathbf{k}, \mathbf{k}_1}. \quad (23)$$

where δ now correspond to the Kronecker delta function.

Construction of bridge processes We are now poised with the task of constructing a conditional probability path with tractable rates, that achieve conditions (21). Since we are dealing with discrete variables, direct interpolation in data space as achieved by equation (14) is impossible. We will follow instead Fitzsimmons et al. (1992) and construct the conditional probability path as a Markov Bridge.

We assume access to a *reference process* characterized by the rate $\mathbf{R}_t(\mathbf{k}, j)$. For this process, we also assume we have closed-form solutions to its corresponding master equation Eq. (19), expressed in terms of the conditional probability $\tilde{q}_{s|t}(\mathbf{k}_1|\mathbf{k})$ of being in state \mathbf{k}_1 at time s when the state was \mathbf{k} at time t . Following Fitzsimmons et al. (1992), we can now construct a Markov bridge that satisfies (21) with a tractable conditional rate given by:

$$\mathbf{W}_t(\mathbf{k}, j|\mathbf{k}_0, \mathbf{k}_1) = \mathbf{R}_t(\mathbf{k}, j) \frac{\tilde{q}_{1|t}(\mathbf{k}_1|\mathbf{k})}{\tilde{q}_{1|t}(\mathbf{k}_1|\mathbf{j})}. \quad (24)$$

Note the slight abuse of notation: the conditional rate is *not* a conditional probability. We also have a corresponding close form for the conditional distribution:

$$q_t(\mathbf{k}|\mathbf{k}_0, \mathbf{k}_1) = \frac{\tilde{q}_{1|t}(\mathbf{k}_1|\mathbf{k}) \tilde{q}_{t|0}(\mathbf{k}|\mathbf{k}_0)}{\tilde{q}_{1|0}(\mathbf{k}_1|\mathbf{k}_0)}. \quad (25)$$

Note, that the rate only depends on the final time condition \mathbf{k}_1 , but not on the initial state \mathbf{k}_0 . This fact is a result of the Markov nature of the reference process. Now since $\mathbf{W}_t(\cdot, \cdot|\mathbf{k}_0, \mathbf{k}_1)$ is known, we can proceed similarly to the flow matching methodology and find the desired target rate trough the marginalization trick as wshown in the next paragraph:

$$\mathbf{W}_t(\mathbf{k}, j) = \mathbb{E}_{\pi_t(\mathbf{k}_0, \mathbf{k}_1|\mathbf{k})} [\mathbf{W}_t(\mathbf{k}, j|\mathbf{k}_0, \mathbf{k}_1)] \quad (26)$$

this is a similar result to equation (16), where now the rate $\mathbf{W}_t(\mathbf{k}, j)$ fulfills the Master equation equation (19) while attaining the target probability flow q_t . The expectation is again over the posterior

probability of the end points, now conditioned on the state \mathbf{k} at time t , obtained with Bayes theorem as:

$$\pi_t(\mathbf{k}_0, \mathbf{k}_1 | \mathbf{k}) = \frac{q_t(\mathbf{k} | \mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{k}_0) \nu(\mathbf{k}_1)}{q_t(\mathbf{k})}. \quad (27)$$

If one is then able to obtain parametric estimate π_t^θ of the posterior π_t (27) from the data, one should be capable of obtaining the target rates by performing the average (26). We can view the posterior learning problem as a probabilistic multi-class classification task. We thus introduce a time-dependent neural network classifier h_t^θ such that the vector of posterior probabilities is given by the *softmax* function $\text{softmax}(h_t^\theta)$. We train the classifier function with the cross-entropy loss, which in this context will be referred to as the Markov jump bridge loss:

$$\mathcal{L}_{\text{MJB}} = \mathbb{E}_{t, (\mathbf{k}_0, \mathbf{k}_1), \mathbf{k}_t} \log h_t^\theta(\mathbf{k}_t, \mathbf{k}_1) \quad (28)$$

where the expectation runs over $t \sim \mathcal{U}(0, 1)$, $(\mathbf{k}_0, \mathbf{k}_1) \sim \mu \times \nu$ and $\mathbf{k}_t \sim q_t(\cdot | \mathbf{k}_0, \mathbf{k}_1)$. Similar to the square loss problem (18), one can show that the minimizer of the cross entropy loss is given by the proper conditional (posterior) probability as one can relate the cross entropy loss to the regression problem [Bishop & Nasrabadi (2006)] (it suffices to see the regression problem over the one hot encoding expression of the likelihood). In the following we impose further conditions on our reference process such as to make our computations more tractable.

Factorizing over Dimensions In our application, we have to model D dimensional data of the form $\mathbf{k} \in \{0, 1, \dots, S\}^D$. Here the state space size grows exponentially with D , which leads to computationally intractable target processes transitions. To avoid this problem we will factorize first the reference process such that all coordinates k_t^d of \mathbf{k}_t are independent processes defined by individual transition rates R_t^d .

$$R_t(\mathbf{k}, \mathbf{j}) = \sum_{d=1}^D R_t^d(k^d, j^d) \prod_{l \neq d} \delta_{j^l, k^l} \quad (29)$$

where for each dimension the marginal probability will follow its own master equation:

$$\frac{\partial \tilde{q}_t^d(k^d)}{\partial t} = \sum_{j^d \neq k^d} R_t^d(k^d, j^d) \tilde{q}_t^d(j^d) - \sum_{j^d \neq k^d} R_t^d(j^d, k^d) \tilde{q}_t^d(k^d) \quad (30)$$

And the full marginal follows $\tilde{q}_t(\mathbf{k}) = \prod_d \tilde{q}_t^d(k^d)$. Notice that due to equation (24) the particular structure of the reference will lead in turn to transition rates for the conditional bridge processes with the following form:

$$W_t(\mathbf{k}, \mathbf{j} | \mathbf{k}_0, \mathbf{k}_1) = \sum_{d=1}^D W_t^d(k^d, j^d | k_1^d) \prod_{l \neq d} \delta_{j^l, k^l} \quad (31)$$

Then again, by the Markovian structure in (24) the expression is independent of the initial state \mathbf{k}_0 . If we apply Eq. (26) to Eq. (31) we see that for rates of the target process only a single component j^d of a vector \mathbf{j} will change. This allows for efficient simulations. That is, one has the form for the target:

$$W_t(\mathbf{k}, \mathbf{j}) = \sum_{d=1}^D W_t^d(k^d, \mathbf{j}) \prod_{l \neq d} \delta_{k^l, j^l} \quad (32)$$

where we have

$$W_t^d(k^d, \mathbf{j}) = \sum_{m=1}^S \pi_t(k_1^d = m | \mathbf{j}) W_t^d(k^d, j^d | k_1^d = m) \quad (33)$$

The required posterior probabilities $\pi_t(k_1^d = m | \mathbf{j})$ for the *individual* coordinates of the end states \mathbf{k}_1 given that the state $\mathbf{k}_t = \mathbf{j}$ are much more efficient to be learned and approximated by neural networks compared to the full joint probability $\pi_t(\mathbf{k}_1 | \mathbf{j})$ required in (26). We now proceed to introduce a reference process with a close form solution to the master equation.

Multivariate Random Telegraph process We now introduce a simple reference process that leads to close form analytical solutions for \tilde{q} . This process is an $S + 1$ -state generalization of the Telegraph process for binary systems, typically used to model burst noise in semi-conductors or bit-flips in communication channels. We assume that the transition probability from all other states to a state k is uniform and described by a rate function $\beta_t := R_t^d(k^d, j^d)$ for all k and j , leading to the following Master equation for $t > s$:

$$\partial_t \tilde{q}_{t|s}(n|m) = \beta(t)(1 - S\tilde{q}_{t|s}(n|m)), \quad (34)$$

where for clarity we have omitted d . Here we focus on the the conditional distribution, as this equation corresponds to a master equation where we enforce impose condition $\tilde{q}_{s|s}(n|m) = \delta_{m,n}$. In the expression (34) we have use the fact that $\sum_{l=1}^S q_{t|s}(l|m) = 1$. This linear, first-order differential equation is solved by

$$\tilde{q}_{t|s}(n|m) = 1/S + w_{t,s} (-1/S + \delta_{m,n}), \quad (35)$$

where

$$w_{t,s} := \exp \left(-S \int_s^t \beta(r) dr \right). \quad (36)$$

In this work we will assume a constant rate function $\beta(t) = \beta$ with hyperparameter $\beta > 0$. Using Eq.(24), we obtain for dimension d and $k \neq j$:

$$W_t^d(k^d = k, j^d = j | k_1) = 1 + \frac{w_{1,t}S}{1 - w_{1,t}} \delta_{k_1,k} + w_{1,t} \delta_{k_1,j}, \quad (37)$$

the equality follows from the binary nature of the Kronecker delta variables, i.e., $\delta_{k_1,k} \in \{0, 1\}$. With the known expressions for the conditional rate in Eq. (37), we can now compute the averages over the posterior in Eq. (26) to obtain W_t^d . This is derived from the posterior expectation as

$$\begin{aligned} W_t^d(k^d, \mathbf{j}) &= \sum_{k_1=1}^S \pi_t(k_1|\mathbf{j}) W_t^d(k^d, j^d | k_1) \\ &= 1 + \frac{w_{1,t}S}{1 - w_{1,t}} \pi_t(k^d|\mathbf{j}) + w_{1,t} \pi_t(j^d|\mathbf{j}), \end{aligned} \quad (38)$$

where the last equality is obtained by substituting Eq. 37. To obtain the target rate W_t , one must learn to approximate the posterior in Eq. 27, and from there, apply Eq. 38. Learning the posterior probabilities $q_t(\mathbf{k}_0, \mathbf{k}_1 | \mathbf{k})$ will involve drawing a large number of random samples $\mathbf{k}_0, \mathbf{k}_1, \mathbf{k}_t$ from their joint probability according to the Markov bridge.

B.3 Proofs

Continuity Equation Here we show how one can construct the marginal vector field from an average of the conditional vector field:

$$\partial_t p_t(\mathbf{x}) \stackrel{(i)}{=} \int \partial_t p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) \mu(\mathbf{x}_0) \nu(\mathbf{x}_1) d\mathbf{x}_0 d\mathbf{x}_1 \quad (39)$$

$$\stackrel{(ii)}{=} - \int \nabla \cdot [u_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1)] \mu(\mathbf{x}_0) \nu(\mathbf{x}_1) d\mathbf{x}_0 d\mathbf{x}_1 \quad (40)$$

$$\stackrel{(iii)}{=} - \int \nabla \cdot [u_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) \rho(\mathbf{x}_0, \mathbf{x}_1 | \mathbf{x}) p_t(\mathbf{x})] d\mathbf{x}_0 d\mathbf{x}_1 \quad (41)$$

$$\stackrel{(iv)}{=} - \nabla \cdot \left[p_t(\mathbf{x}) \int u_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) \rho(\mathbf{x}_0, \mathbf{x}_1 | \mathbf{x}) d\mathbf{x}_0 d\mathbf{x}_1 \right] \quad (42)$$

$$\stackrel{(v)}{=} - \nabla \cdot [u_t(\mathbf{x}) p_t(\mathbf{x})], \quad (43)$$

where we have use in (iii) bayes rule for the posterior:

$$\rho_t(\mathbf{x}_0, \mathbf{x}_1 | \mathbf{x}) = \frac{p_t(\mathbf{x}|\mathbf{x}_0, \mathbf{x}_1) \mu(\mathbf{x}_0) \nu(\mathbf{x}_1)}{p_t(\mathbf{x})}. \quad (44)$$

Master Equation Here we show how one can construct the target rate from an average of the conditional rate:

$$\frac{\partial}{\partial t} q_t(\mathbf{k}) = \frac{\partial}{\partial t} \sum_{\mathbf{k}_0, \mathbf{k}_1} q_t(\mathbf{k} | \mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{k}_0) \nu(\mathbf{k}_1) \quad (45)$$

$$= \sum_{\mathbf{k}_0, \mathbf{k}_1} \frac{\partial}{\partial t} q_t(\mathbf{k} | \mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{k}_0) \nu(\mathbf{k}_1) \quad (46)$$

$$= \sum_{\mathbf{k}_0, \mathbf{k}_1} \sum_{\mathbf{j}} W_t(\mathbf{k}, \mathbf{j} | \mathbf{k}_1) q_t(\mathbf{j} | \mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{k}_0) \nu(\mathbf{k}_1) \quad (47)$$

$$\stackrel{(i)}{=} \sum_{\mathbf{j}} \left\{ \sum_{\mathbf{k}_1, \mathbf{k}_0} W_t(\mathbf{k}, \mathbf{j} | \mathbf{k}_1) \pi_t(\mathbf{k}_1, \mathbf{k}_0 | \mathbf{j}) \right\} q_t(\mathbf{j}) \quad (48)$$

$$= \sum_j W_t(\mathbf{k}, \mathbf{j}) q_t(\mathbf{j}) \quad (49)$$

where we have used the posterior

$$\pi_t(\mathbf{k}_0, \mathbf{k}_1 | \mathbf{j}) = \frac{q_t(\mathbf{j} | \mathbf{k}_0, \mathbf{k}_1) \mu(\mathbf{k}_0) \nu(\mathbf{k}_1)}{q_t(\mathbf{j})}. \quad (50)$$

Conditional Rate We now obtain the rate of the reference process conditioned on the end points, that is a jump process bridge $P(\cdot | \mathbf{k}_1, \mathbf{k}_0)$. We omit the dimension index d for clarity.

$$W_t(\mathbf{k}, \mathbf{j} | \mathbf{k}_1) = \lim_{\Delta t \rightarrow 0} \left[\frac{\tilde{q}_{t+\Delta t | t}(\mathbf{k} | \mathbf{j}, \mathbf{k}_1) - \delta_{\mathbf{k}, \mathbf{j}}}{\Delta t} \right] \quad (51)$$

$$= \lim_{\Delta t \rightarrow 0} \left[\frac{\tilde{q}_{1, t+\Delta t, t}(\mathbf{k}_1, \mathbf{k}, \mathbf{j}) - \delta_{\mathbf{k}, \mathbf{j}}}{\Delta t \tilde{q}_{1, t}(\mathbf{k}_1, \mathbf{j})} \right] \quad (52)$$

$$= \lim_{\Delta t \rightarrow 0} \left[\frac{\tilde{q}_{1|t+\Delta t}(\mathbf{k}_1 | \mathbf{k}) \tilde{q}_{t+\Delta t | t}(\mathbf{k} | \mathbf{j}) \tilde{q}_t(\mathbf{j})}{\Delta t \tilde{q}_{1|t}(\mathbf{k}_1 | \mathbf{j}) \tilde{q}_t(\mathbf{j})} - \frac{\tilde{q}_{1|t+\Delta t}(\mathbf{k}_1 | \mathbf{k}) \delta_{\mathbf{k}, \mathbf{j}}}{\tilde{q}_{1|t}(\mathbf{k}_1 | \mathbf{j}) \Delta t} \right] \quad (53)$$

$$= \lim_{\Delta t \rightarrow 0} \left[\frac{\tilde{q}_{1|t+\Delta t}(\mathbf{k}_1 | \mathbf{k})}{\tilde{q}_{1|t}(\mathbf{k}_1 | \mathbf{j})} \left(\frac{\tilde{q}_{t+\Delta t}(\mathbf{k} | \mathbf{j}) - \delta_{\mathbf{k}, \mathbf{j}}}{\Delta t} \right) \right] \quad (54)$$

$$= \frac{\tilde{q}_{1|t}(\mathbf{k}_1 | \mathbf{k})}{\tilde{q}_{1|t}(\mathbf{k}_1 | \mathbf{j})} R_t(\mathbf{k}; \mathbf{j}) \quad (55)$$

Conditional Probability/ Markov Bridge Here we obtain the expression for Eq. (25), this equation holds for any Markov process

$$\begin{aligned} p_t(\mathbf{x} | \mathbf{x}_0, \mathbf{x}_1) &\stackrel{(i)}{=} \frac{p(\mathbf{x}_0, \mathbf{x}_t = \mathbf{x}, \mathbf{x}_1)}{p(\mathbf{x}_0, \mathbf{x}_1)} \\ &\stackrel{(ii)}{=} \frac{p(\mathbf{x}_1 | \mathbf{x}_t = \mathbf{x}) p(\mathbf{x}_t = \mathbf{x} | \mathbf{x}_0) p(\mathbf{x}_0)}{p(\mathbf{x}_0, \mathbf{x}_1)} \\ &\stackrel{(iii)}{=} \frac{p_{1|t}(\mathbf{x} | \mathbf{x}) p_{t|0}(\mathbf{x} | \mathbf{x}_0) p(\mathbf{x}_0)}{p_{1|0}(\mathbf{x}_1 | \mathbf{x}_0) p(\mathbf{x}_0)} \\ &= \frac{p_{1|t}(\mathbf{x}_1 | \mathbf{x}) p_{t|0}(\mathbf{x} | \mathbf{x}_0)}{p_{1|0}(\mathbf{x}_1 | \mathbf{x}_0)} \end{aligned}$$

Here as before one only requires the Markovianity assumption of p , which means that for $t > s$ one can write $p(\mathbf{x}_t, \mathbf{x}_s) = p_{t|s}(\mathbf{x}_t | \mathbf{x}_s) p(\mathbf{x}_s)$ as well as Bayes' rule.

C Architecture

Mode embeddings To effectively model the multimodal nature of the data, we embed independently each input mode, x_t^d and k_t^d , into a high-dimensional vector space $\mathbb{R}^{h_{\text{emb}}}$ using Multi-layer perceptrons

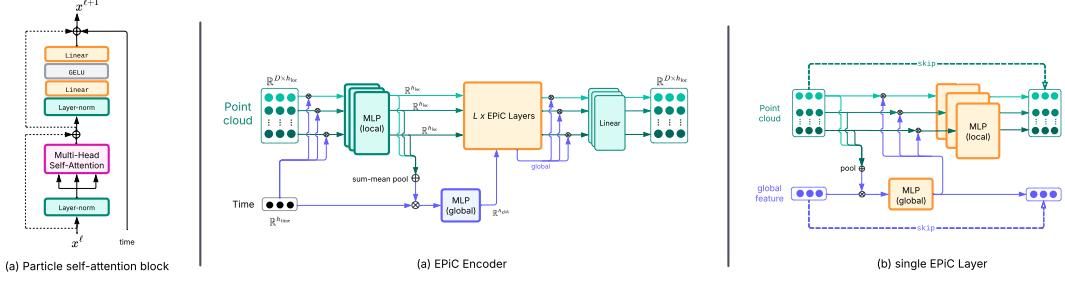


Figure 4: Detail of the particle self-attention block used for our MMF model and the EPiC encoder used for our baseline model.

(MLP) consisting of two linear layers with a GELU non-linear activation function in between. To process the flavor tokens k_t^d we replace the first layer of the MLP with a learnable lookup table implemented in PyTorch via `nn.Embedding`. The time variable $t \in [0, 1]$ is encoded into $t_{\text{emb}} \in \mathbb{R}^{h_{\text{emb}}}$ using a Fourier feature (FF) embedding of Tancik et al. (2020). The resulting embeddings for each mode are then summed with the embedded time vector $t_{\text{emb}} = \text{FF}(t)$ to form hidden kinematic and flavor representations:

$$\mathbf{x}'_t = \text{MLP}(\mathbf{x}_t) + t_{\text{emb}}, \quad \mathbf{k}'_t = \text{MLP}(\mathbf{k}_t) + t_{\text{emb}}. \quad (56)$$

These are subsequently fed into a time-dependent encoder.

Multimodal particle transformer Given that the dataset consists of particle clouds, where the ordering of constituents is unimportant, it is essential to approximate the generators of the dynamical process with a permutation-equivariant architecture. In this work we employ a particle transformer architecture, or *ParticleFormer* for short, as the core component of our multimodal encoder. The ParticleFormer processes the embedded time variable along with the embedded particle-level features with a stack of multi-head self-attention blocks. Details are shown in Fig. 4 (a). We use a generic GPT-style self-attention block without causal masking and positional encoding to guarantee permutation equivariance. Particle transformers of this sort were first used for jet tagging on simulated data by Qu et al. (2022), producing state-of-the-art results when compared to previous methods such as GANs and graph neural networks.

Our multimodal encoder consists of two mode-specific encoders that feed into a *fused* encoder:

$$F_{\text{kin}} = \text{ParticleFormer}_{L_1}(t_{\text{emb}}, \mathbf{x}'_t), \quad (57)$$

$$F_{\text{flav}} = \text{ParticleFormer}_{L_2}(t_{\text{emb}}, \mathbf{k}'_t), \quad (58)$$

$$F_{\text{fuse}} = \text{ParticleFormer}_L(t_{\text{emb}}, F_{\text{kin}} \otimes F_{\text{flav}}), \quad (59)$$

Here, \otimes denotes feature concatenation, and the parameters L , L_1 , and L_2 represent the number of self-attention layers in each encoder. This architecture provides flexibility, enabling the network to capture intra-modal correlations in the early mode-specific encoders, while cross-modal correlations between kinematics and flavor are learned in the subsequent fused encoder. Notice here that this particular choice of first encoding each modality and then fusing them is somewhat arbitrary. Other multimodal frameworks, like using cross-attention between modes could be implemented or added to our framework. Exploring other multimodal setups is left for future studies.

The output state of the fused encoder is split into two equal-sized states $F_{\text{fuse}} = H_{\text{kin}} \otimes H_{\text{flav}}$, which are then processed by the (continuous) regressor and the (discrete) classifier heads. Before passing these to each head, we add the mode-specific residual states and the time embedding:

$$H'_{\text{kin}} = H_{\text{kin}} + F_{\text{kin}} + t_{\text{emb}}, \quad H'_{\text{flav}} = H_{\text{flav}} + F_{\text{flav}} + t_{\text{emb}}. \quad (60)$$

The heads consist of two-layered MLPs with a GELU activation function in between that output the velocity field and the posterior classifier, respectively:

$$u_t^\theta \otimes h_t^\theta = \text{MLP}(H'_{\text{kin}}) \otimes \text{MLP}(H'_{\text{flav}}). \quad (61)$$

Uncertainty network As discussed in the text, we promote the uncertainty weights of our multi-modal loss (8) to time-dependent functions parametrized with a neural network. For this network we use a single Fourier feature layer with a 128-dimensional hidden state followed by a linear projection layer $(w_t^1, w_t^2) = \text{Linear}(\text{FF}(t))$. This module is only used during the training phase.

D Sampling algorithm details

To generate the particle kinematics we directly solve the ODE (10) using any well-known integration method. For simplicity, we integrate using Euler’s first-order method:

$$x_{t+\Delta t}^d = x_t^d + u_t^{\theta, d}(x_t^d, k_t^d) \Delta t \quad (62)$$

where Δt is a small time-step, and $u_t^{\theta, d}$ is the parametric velocity field for each particle.

τ -leaping To efficiently simulate the random telegraph process, we employ τ -leaping. Rather than resolving each individual transition sequentially, tau-leaping assumes the total number of per-particle transitions Δn_m^d into the flavor token m , occurring within a small time window Δt , follows a *Poisson distribution*,

$$\Delta n_m^d \sim \text{Poisson}(W_t^{\theta, d}(k_{t+\Delta t}^d = m, k_t^d, x_t^d) \Delta t). \quad (63)$$

This approximation holds under the assumption that individual jumps within Δt occur independently and with probabilities proportional to $W_t^{\theta, d} \Delta t$. In this regime, the total number of jumps Δn_j^d can be understood as arising from many independent Bernoulli trials, which, in the limit of small probabilities, naturally follows a Poisson distribution. By appropriately selecting Δt , tau-leaping provides a useful trade-off between accuracy and computational efficiency, enabling the effective simulation of discrete jumps without the need for explicitly resolving each transition at every infinitesimal time step. Explicitly, at each time step, the discrete flavor state for each particle is updated via

$$k_{t+\Delta t}^d = \left[k_t^d + \sum_{m=1}^S (m - k_t^d) \Delta n_m^d \right] \bmod S. \quad (64)$$

Here, we take the modulo of the vocabulary size S to deal with cases where the updated state inside the bracket results in an integer outside of $\mathcal{F} = \{0, \dots, 7\}$. An alternative is to clamp the output to the boundary values so they remain within \mathcal{F} , with the expense of biasing the generation towards these tokens (in our case these correspond to the photon and the anti-muon).

The combined sampling procedure for the hybrid states consists of iteratively updating the continuous and discrete features using Euler steps (62) for the kinematics and tau-leaping steps (64) for the flavor tokens.

Temperature scaling As discussed in Sec. 2, the training dataset exhibits a pronounced class imbalance, with jets containing much more photons and charged hadrons compared to neutral hadrons and leptons. Such imbalances are well known to hinder classification performance [Johnson & Khoshgoftaar (2019)]. A common way to alleviate their impact is to recalibrate the posterior probabilities (61) through *temperature scaling* [Guo et al. (2017)]. Specifically, we introduce a temperature hyperparameter T as

$$\pi_t^\theta = \text{softmax}\left(\frac{\text{MLP}(H'_{\text{flav}})}{T}\right). \quad (65)$$

This rescaling of the logits is only applied during generation, leaving training unaffected. Larger values $T > 1$ soften the logits, reducing the relative differences between classes and approaching a uniform distribution as $T \rightarrow \infty$. Conversely, smaller values $T < 1$ sharpen the distribution, accentuating class differences and approaching a one-hot assignment in the limit $T \rightarrow 0$. Temperature scaling has also been widely used in natural language processing and classification tasks, where it is known to balance class probabilities and improve calibration. In our experiments, we investigate the impact of the temperature on generation quality.

Particle multiplicities A fundamental limitation of dynamics-based flow models comes from their inability to handle particle-clouds with varying number of particles. Despite masking zero-padded entries during training, the generation step requires explicit conditioning on the number of particles per jet. This constraint originates from the continuity equation (9) governing the reference dynamics, which enforces particle-number conservation along each trajectory, thereby precluding the spontaneous creation or annihilation of particles during the evolution. To address this limitation and allow variable particle numbers within our generative framework we fit a 150-dimensional categorical distribution to the empirical particle multiplicity distribution. During generation, for each jet we sample the particle multiplicity N from this auxiliary model, and subsequently apply the corresponding mask (N ones followed by $150 - N$ zeros) to the initial source data.

E Experiment details

EPiC-FM baseline We train the baseline on the same target AOJ datasets. However, since flow-matching can only handle continuous variables, the flavor token of each particle is one-hot encoded into unit vectors in \mathbb{R}^S representing flavor assignment probabilities. For the target jets, these probabilities are concatenated with the particle kinematics, forming an augmented continuous feature vector $x_1^d \in \mathbb{R}^{3+S}$. We generate source point-clouds by drawing each point from a standard Gaussian over \mathbb{R}^{3+S} . After generation, to ensure each particle has a unique flavor assignment, we apply an `argmax` operation to the generated assignment probabilities and tokenize back to \mathcal{F} . This strategy has been successfully employed in previous works [Birk et al. (2023); Araz et al. (2024)] for jet and event datasets. For our experiments, we implement the EPiC-FM encoder described in Buhmann et al. (2023a) depicted in Fig. 4 (b) with the following setup: $n_{\text{layers}} = 16$ EPiC layers with $h_{\text{loc}} = 256$ and $h_{\text{glob}} = 16$ for the local and global hidden dimensions. The resulting model has around 5.9 million parameters. Optimization is perform with the Adam algorithm [Kingma & Ba (2014)] with an effective batch size of 256 jets for a maximum of 1500 epochs. A cosine-annealing learning rate schedule is applied for the first 1000 epochs, decaying from 5×10^{-4} to 10^{-5} , followed by 500 epochs at a fixed learning rate of 10^{-5} .

MMF training details We set the gaussian smearing hyperparameter of the flow-matching component to $\sigma = 10^{-5}$ and use a constant stochasticity parameter with $\beta = 0.075$ for the multivariate telegraph process. During training, the time parameter t is sampled uniformly from a slightly reduced unit interval $[\epsilon, 1 - \epsilon]$ with $\epsilon = 10^{-5}$, to prevent numerical instabilities caused at the time boundaries $t = 0, 1$ for the MJB model. We parameterize the combined generators $u_t^\theta \otimes \pi_t^\theta$ using the multimodal architecture introduced in Sec. 3 with a *mid-fusion* setup $(L_1, L_2, L) = (5, 5, 6)$. This configuration balances the intra-modal and cross-modal correlations between kinematics and flavor tokens in separate sub-modules with similar sizes. We fix the number of attention heads and hidden dimensions to $n_{\text{heads}} = 4$, $n_{\text{embd}} = 256$, and $n_{\text{inner}} = 512$, resulting in a model with approximately 5.6 million trainable parameters. Training is performed with the uncertainty weighted loss of Eq. (8). We use an uncertainty network with a 128-dimensional hidden state. Optimization is perform with the Adam algorithm [Kingma & Ba (2014)] with an effective batch size of 256 jets for a maximum of 1500 epochs. A cosine-annealing learning rate schedule is applied for the first 1000 epochs, decaying from 5×10^{-4} to 10^{-5} , followed by 500 epochs at a fixed learning rate of 10^{-5} .

The best models are chosen according to the lowest validation loss. All experiments are run on 16 NVIDIA A100 GPUs.