JET FRAGMENTATION VIA MACHINE LEARNING

Oral Exam

Tony Menzo

University of Cincinnati

May 10, 2022

Monte Carlo Event Generators

- Currently there is no phenomenological model of hadronization which correctly reproduces all experimental data.
- No ‘new’ models of fragmentation in $\sim$ 30 years

Event generators are ubiquitous in HEP:

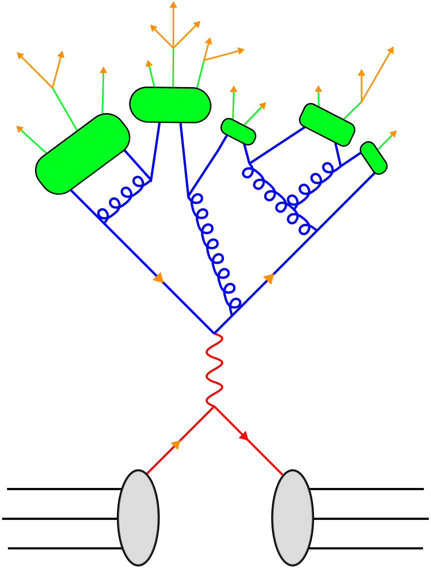
Citation data according to INSPIRE

Event generator	hep-ex (Citations)	hep-ph (Citations)
PYTHIA 6	8,314	4,459
PYTHIA 8	4,528	2,100
HERWIG 6	2,499	1,103
HERWIG++	1,906	971
SHERPA	2,569	1,073

Simulation of hard process

The simulation is ordered according to the magnitude of momentum transfer:

1. Hard interaction
2. Parton Showers
3. Hadronization
4. Unstable particle decay



QCD, Confinement, and Strings

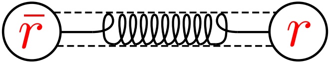
Around Λ_{QCD} particles charged under QCD begin to confine into color neutral hadrons. Separations > 1 fm generate thin $\mathcal{O}(1/\Lambda_{\text{QCD}})$ color flux tubes between charged particles.

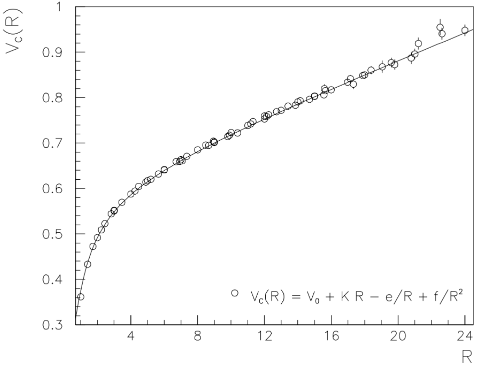
The $q\bar{q}$ potential is given by

$$V_{\text{QCD}}(r) \approx -\frac{4}{3}\frac{\alpha_s}{r} + \kappa r$$

with $\kappa \approx 1$ GeV/fm

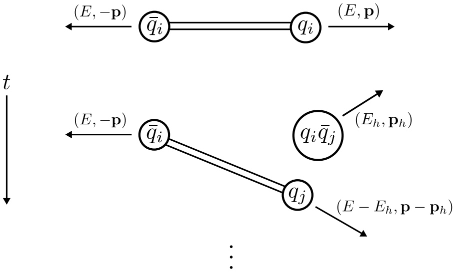
- Motivates a classical description at large separations

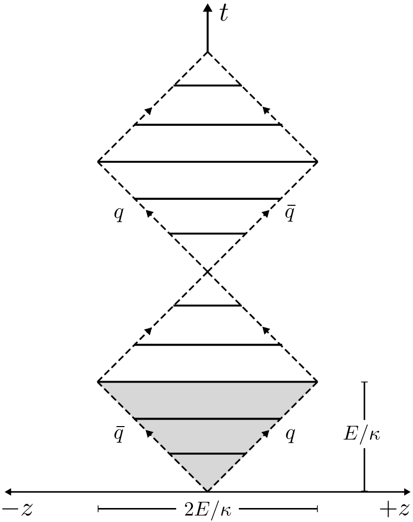




Confirmation from quenched lattice QCD simulation. arXiv:hep-lat/9210003

Lund Model: ***Classical*** model of hadronization utilizing string picture. Assume $r \gg 1 \rightarrow V_{\text{Lund}} = \kappa r$



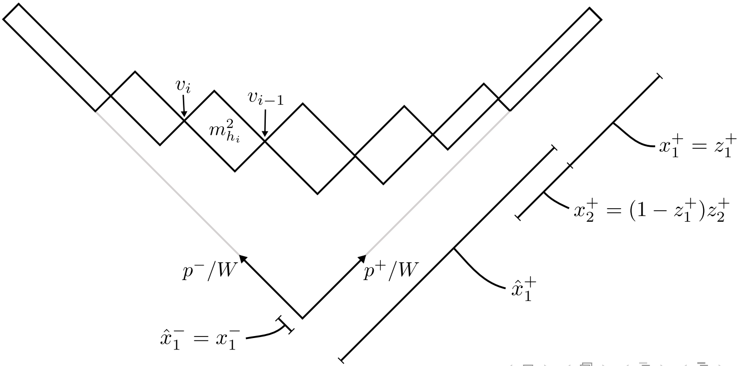


Yo-yo string motion in the absence of string breaks as seen in the center of mass frame of the string system.

• Fragmentation is implemented as a stochastic process through production vertices in momentum space.
• Two inputs: mass of hadron m_h and longitudinal momentum fraction z. Left-right symmetric Lund fragmentation function:

$$f(z)dz \propto \frac{(1-z)^a}{z} \exp\left(-b\frac{m_h^2}{z}\right) dz$$

where a and b are fit parameters.



Particle	Multiplicity	HERWIG 5.9	JETSET 7.4	UCLA 7.4	Expts
Charged	20.96(18)	20.95	20.95	20.88	ADLMO
$\pi^{\pm}$	17.06(24)	17.41	16.95	17.04	ADO
π^0	9.43(38)	9.97	9.59	9.61	ADLO
$\boldsymbol{\eta}$	0.99(4)	1.02	1.00	0.78	ALO
$\rho(770)^0$	1.24(10)	1.18	1.50	1.17	AD
$\omega(782)$	1.09(9)	1.17	1.35	1.01	ALO
$\boldsymbol{\eta'(958)}$	0.159(26)	0.097	0.155	0.121	ALO
f_0**(980)**	0.155(8)	0.111	~0.1	—	ADO
$\mathrm{a}_0(980)^{\pm}$	0.14(6)	0.240	—	—	O
$\phi(1020)$	0.097(7)	0.104	0.194	0.132	ADO
f_2**(1270)**	0.188(14)	0.186	~ 0.2	—	ADO
$\mathrm{f}'_2(1525)$	0.012(6)	0.021	—	—	D
$\mathrm{K}^{\pm}$	2.26(6)	2.16	2.30	2.24	ADO
$\mathrm{K}^{\mathbf{0}}$	2.074(14)	2.05	2.07	2.06	ADLO
$\mathrm{K}^*(892)^{\pm}$	0.718(44)	0.670	1.10	0.779	ADO
$\mathrm{K}^*(892)^0$	0.759(32)	0.676	1.10	0.760	ADO
$\mathrm{K}_2^*(1430)^0$	0.084(40)	0.111	—	—	DO
$\mathrm{D}^{\pm}$	0.187(14)	0.276	0.174	0.196	ADO
D^0	0.462(26)	0.506	0.490	0.497	ADO
$\mathrm{D}^*(2010)^{\pm}$	0.181(10)	0.161	0.242	0.227	ADO
$\mathrm{D}_s^{\pm}$	0.131(20)	0.115	0.129	0.130	O
B^*	0.28(3)	0.201	0.260	0.254	D
$\mathrm{B}_{u,d}^{**}$	0.118(24)	0.013	—	—	D
J/ψ	0.0054(4)	0.0018	0.0050	0.0050	ADLO
$\psi(3685)$	0.0023(5)	0.0009	0.0019	0.0019	DO
χ_{c1}	0.0086(27)	0.0001	—	—	DL
p	1.04(4)	0.863	1.19	1.09	ADO
Δ^{++}	0.079(15)	0.156	0.189	0.139	D
	0.22(6)	0.156	0.189	0.139	O
$\boldsymbol{\Lambda}$	0.399(8)	0.387	0.385	0.382	ADLO
$\boldsymbol{\Lambda}$**(1520)**	0.0229(25)	—	—	—	DO
$\Sigma^{\pm}$	0.174(16)	0.154	0.140	0.118	DO
Σ^0	0.074(9)	0.068	0.073	0.074	ADO
$\Sigma^{\star\pm}$	0.0474(44)	0.111	0.074	0.074	ADO
Ξ^-	0.0265(9)	0.0493	0.0271	0.0220	ADO
$\Xi(1530)^0$	0.0058(10)	0.0205	0.0053	0.0081	ADO
Ω^-	0.0012(2)	0.0056	0.00072	0.0011	ADO
Λ_c^+	0.078(17)	0.0123	0.059	0.026	O

← Meson and baryon yields in Z^0 decays. Underlined: Disagreement with data by more than 3σ. [arXiv:hep-ph/9912292]

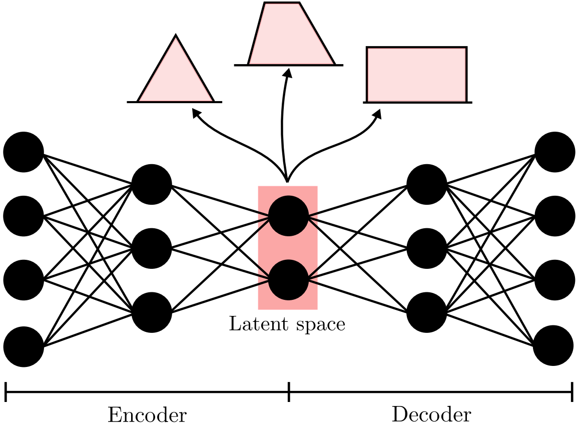
Issues with the Lund model:

1. Many parameters
2. Problems describing baryon production
3. Dependent on analytic $f(z)$

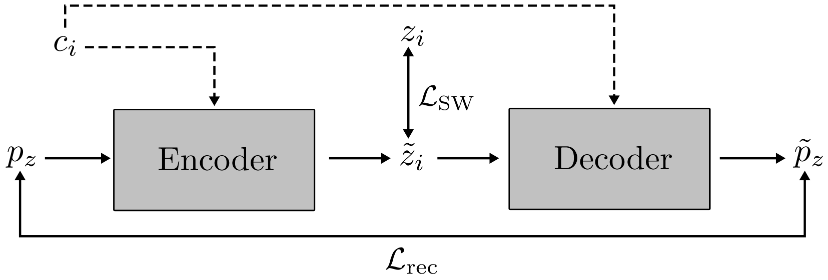
- New paradigm: build phenomenological models which sample directly from empirical distributions.

Sliced-Wasserstein Autoencoder (SWAE)

Goal: Input number sampled from a 'simple' probability distribution and return a number distributed according to the hadronic kinematic distribution

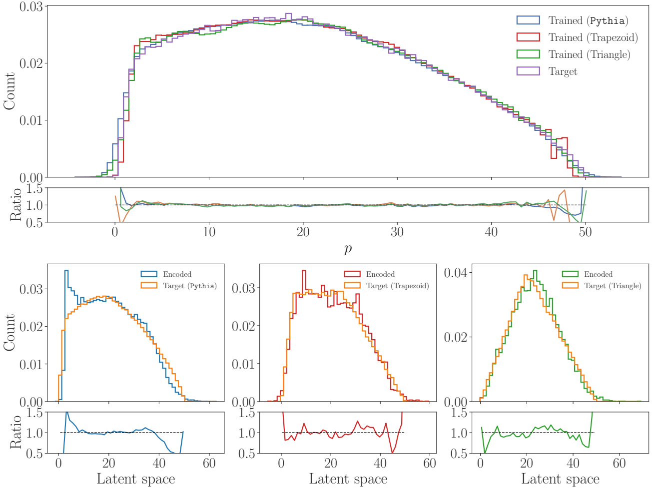


Conditional SWAE



Program architecture for our implementation of the SWAE.

SWAE-trained models



Inclusive first emission magnitude of momentum distribution compared with encoder/decoder results from three SWAE-trained models.

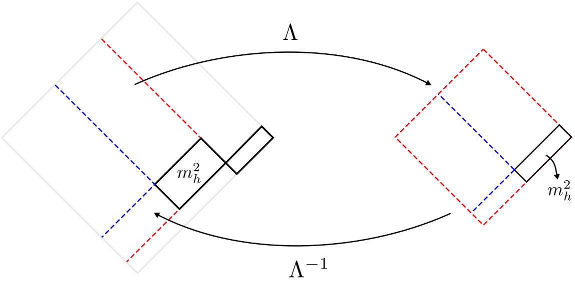
Modified Lund model

We require a change of inputs $(m_h, z) \to (E_{\mathrm{CM}}, \tilde{z})$ and rely on:

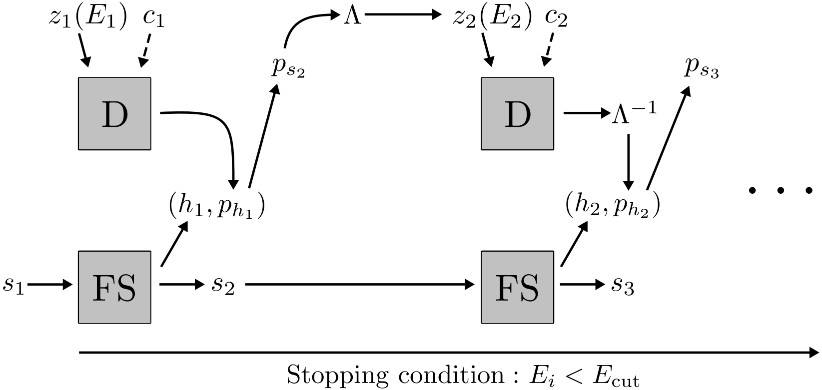
1. Causal disconnection of events

2. Simple rescaling of the kinematic distributions with respect to energy. For example,

$$p'_z = p_{z,\mathrm{ref}} \frac{p_z}{E} \tag{1}$$

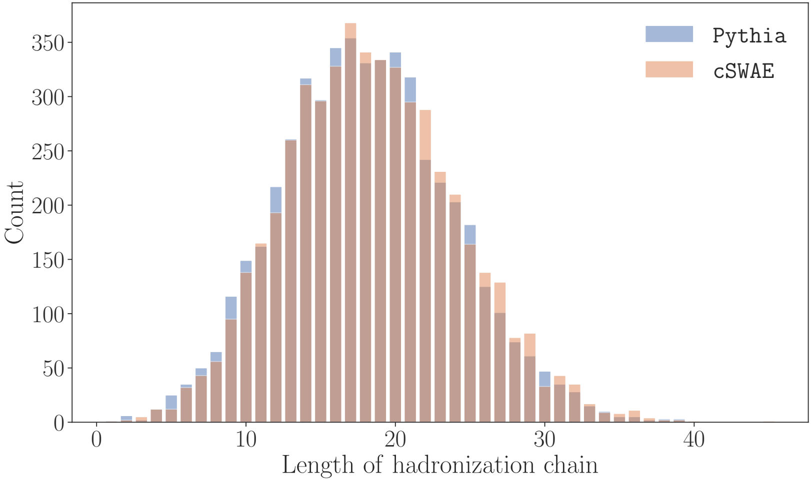


Fragmentation chain



Program architecture for our implementation of the fragmentation chain.

1+1 Dimensional Fragmentation Multiplicities



Simplified hadronization chain, only includes $u(\bar{u})$ and $d(\bar{d})$ quarks as string ends and pion $(\pi^0, \pi^{\pm})$ final states

Generalization to 3+1 Dimensions

• Production pair can now be produced with transverse momenta distributed according to the semi-classical tunneling probability

$$f(p_\perp) \propto \exp(\pi p_\perp^2/\kappa)$$

• Replace $m_h^2 \to m_\perp^2 = m_h^2 + p_\perp^2$

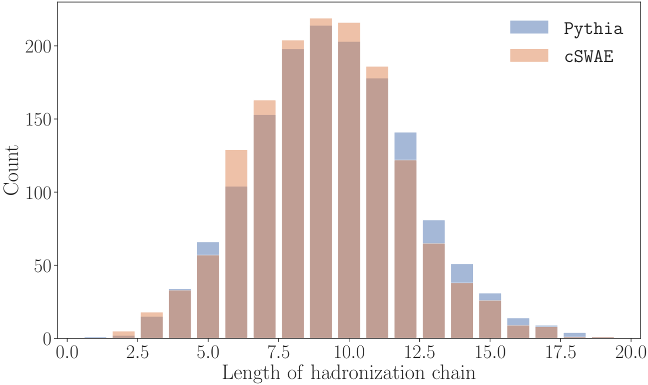
This creates two issues:

1. Our fragmentation chain always samples from the CM frame whereas the transverse momenta are generated in the lab frame.

2. The dependence on $f(z)$ on $m_\perp$ creates a correlation between z and $p_\perp$ through each iteration.

Approximate solution

- Sample p_x and p_y from distribution above
- Train on p_z distributions where the hadron mass is increased by the variance of the $p_\perp$ distribution



- Ideally train on m_T-labeled dataset

Other future directions

1. Generalize to include all quarks and hadronic final states

2. A new model of flavor selection via machine learning

3. Additional event topologies: $gg \to ggh$, $qq\bar{q} \to qq\bar{q}h$, ...

4. Integration with FastJet to test model consequences in jet observables

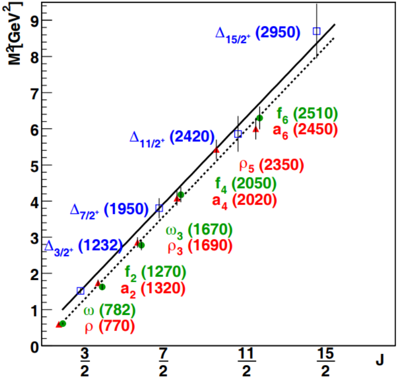
5. Feasibility of collecting training datasets from jet data (CMS Open Data)

Thank you ☺

Back-ups

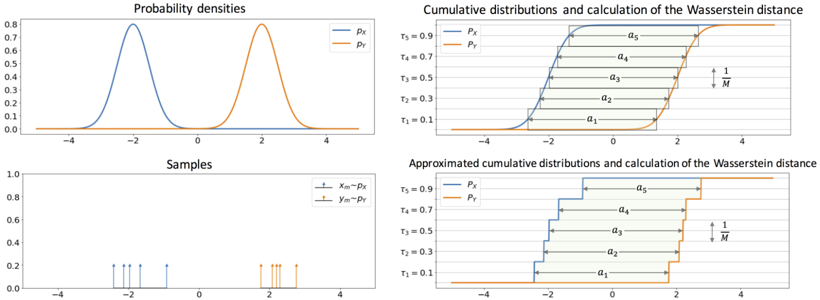
Regge Trajectories

$$J = \alpha_0 + \alpha' M^2$$



Leading Regge trajectories $J = L + S$. Klempt, E., and B. Ch Metsch. "Multiplet classification of light-quark baryons." The European Physical Journal A 48.9 (2012): 1-15.

Wasserstein Distance



Wasserstein distance for 1D PDF. Kolouri, Soheil, et al. "Sliced Wasserstein auto-encoders." International Conference on Learning Representations. 2018.