Flow Annealed Importance Sampling Bootstrap meets Differentiable Particle Physics

NeurIPS Workshop ML4PS

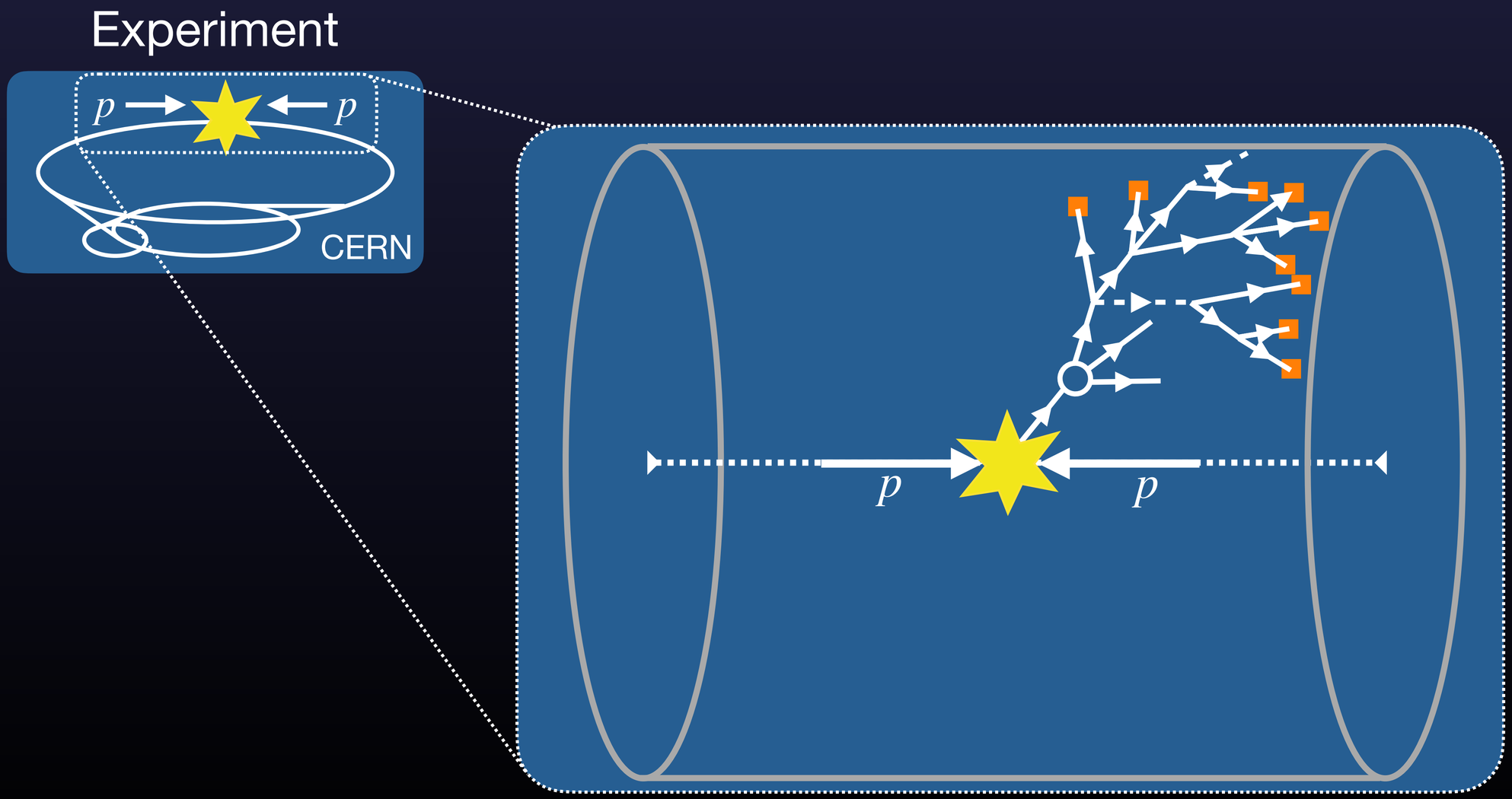




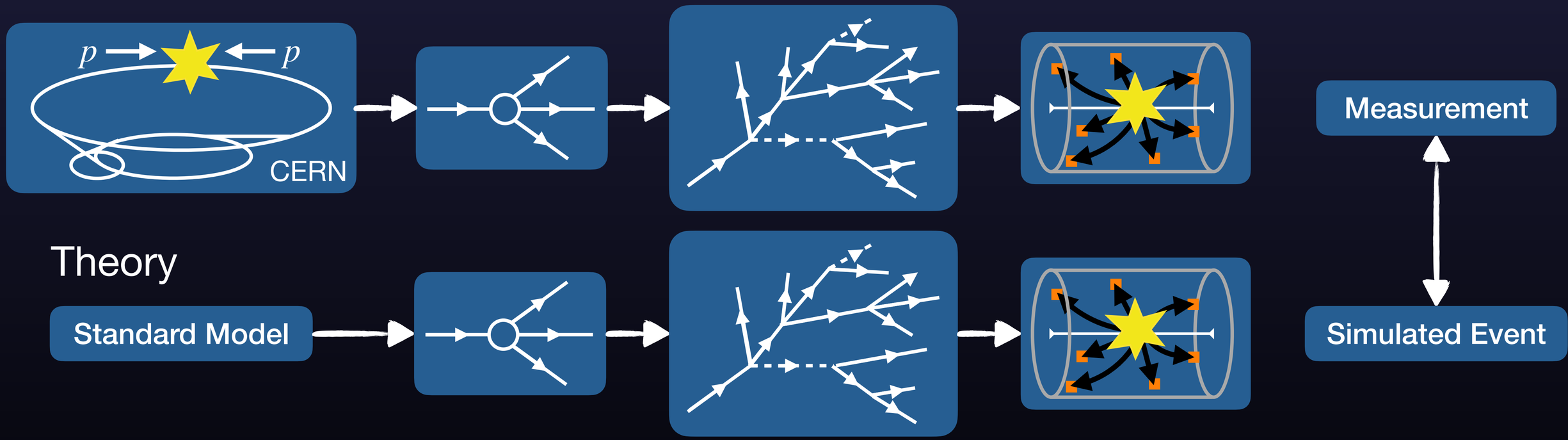
Annalena Kofler[1,2,3] Vincent Stimper[1,4,5] Mikhail Mikhasenko[6,7] Michael Kagan[8] Lukas Heinrich[3]

[1]Max Planck Institute for Intelligent Systems Tübingen, [2]Max Planck Institute for Gravitational Physics Potsdam, [3]Technical University of Munich, [4]Isomorphic Labs, [5]University of Cambridge, [6]University of Bochum, [7]ORIGINS Excellence Cluster Munich, [8]SLAC National Accelerator Laboratory

How can we learn more about particles?



How can we analyze data?



→ Find deviations from theory

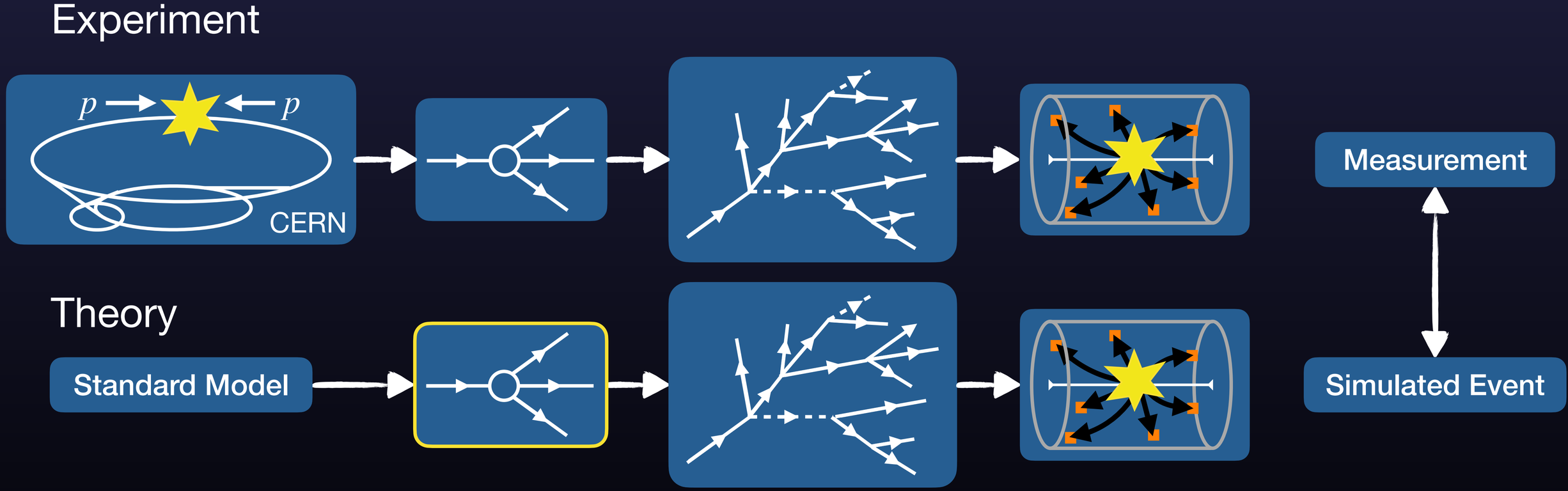
How to improve the precision of measurements?

→ Take more data!

Number of
events
until 2030

ATLAS Software and Computing HL-LHC Roadmap, 2022.

What does this mean for the simulations?

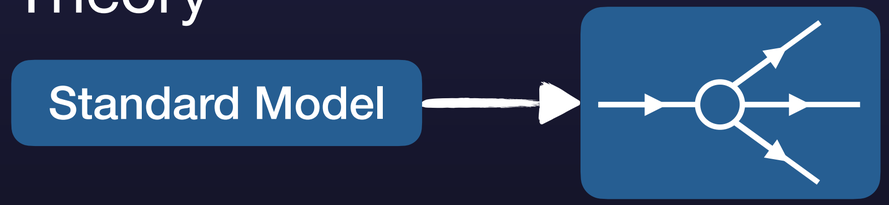


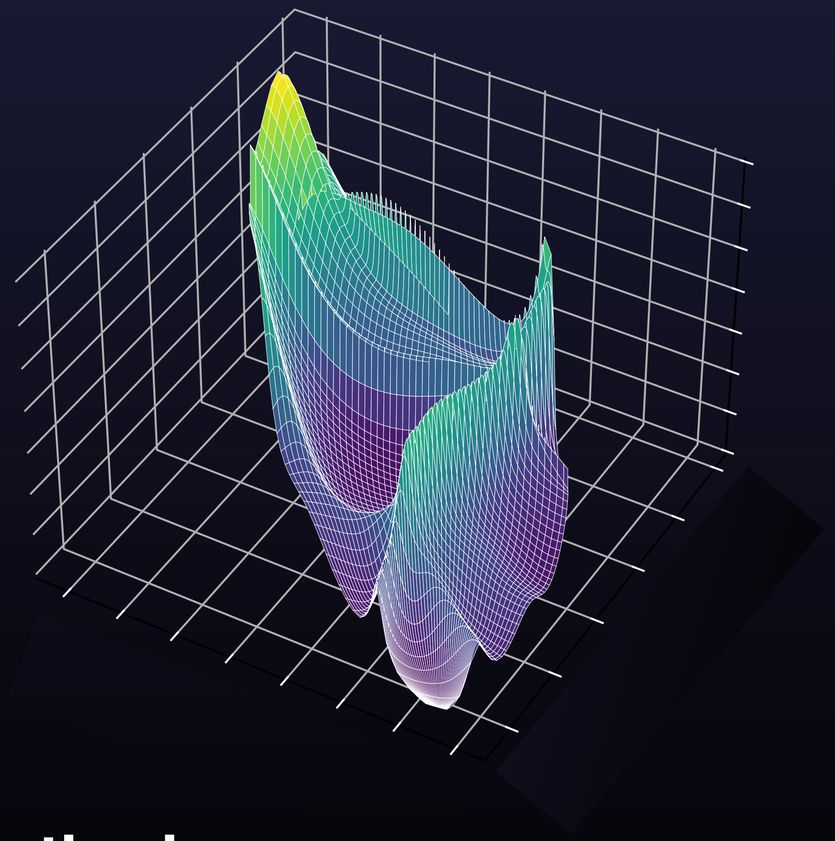
→ We need to speed up the simulation!

→ In this work: Event generation

What is event generation from a ML perspective?

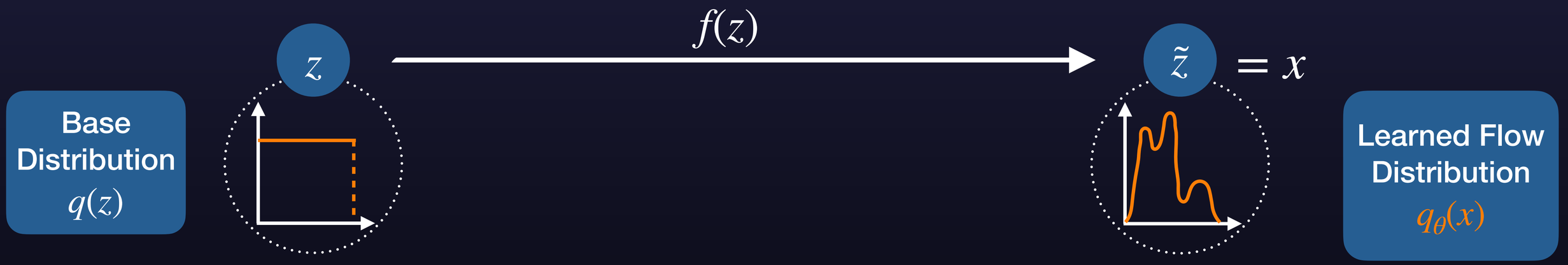
Theory





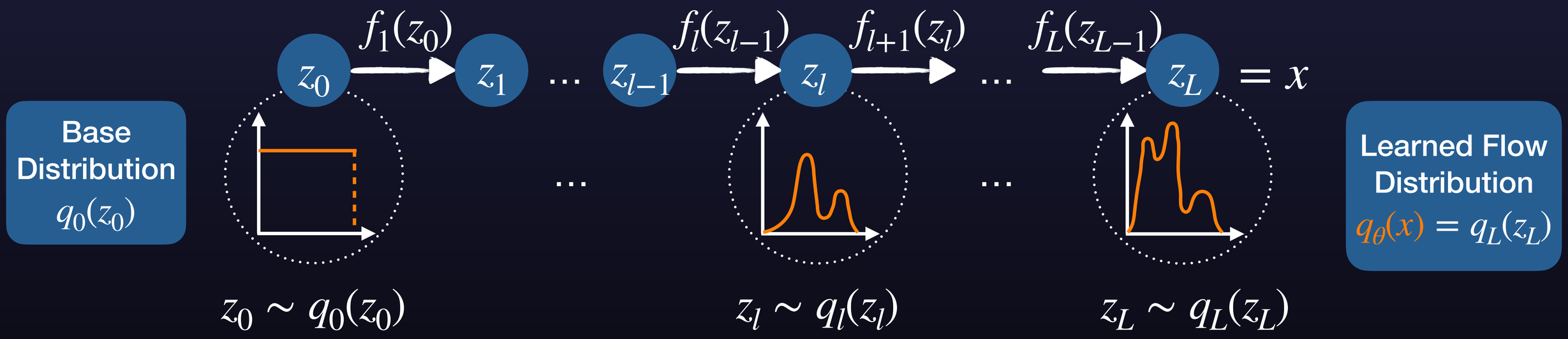
- Analytic formula for unnormalized distribution $p(x)$ (“matrix element”)
- $p(x)$ describes the outgoing particles
- Sample from this distribution
- **Normalizing flow instead of standard methods**

What is a normalizing flow?



Rezende and Mohamed, “Variational Inference with Normalizing Flows.” ICML’15.
Durkan et al., “Neural Spline Flows.”, NeurIPS’19.

What is a normalizing flow?

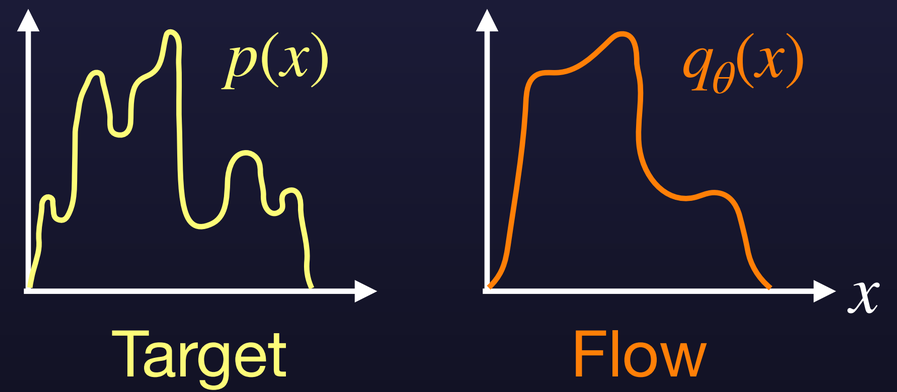


→ Generate samples x and evaluate density $q_\theta(x)$

Rezende and Mohamed, “Variational Inference with Normalizing Flows.” ICML’15.
Durkan et al., “Neural Spline Flows.”, NeurIPS’19.

How to train a normalizing flow?

3 Methods



Samples from target $x \sim p(x)$

(1) Forward KL Divergence (fKLD)

Samples from flow $x \sim q_\theta(x)$ and density evaluation of $p(x)$

(2) Reverse KL Divergence (rKLD)

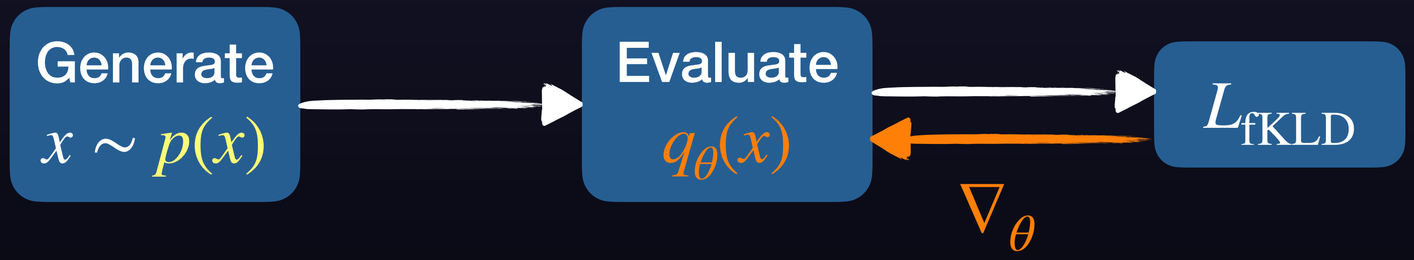
(3) Flow Annealed Importance Sampling Bootstrap (FAB)

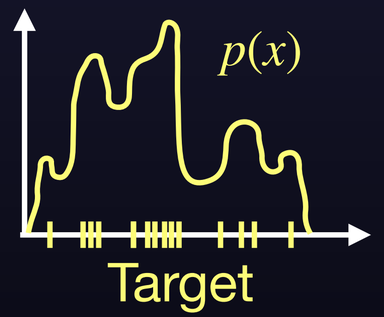
Midgley, Stimper, et al., “Flow Annealed Importance Sampling Bootstrap”, ICLR, 2023.

Training with Samples

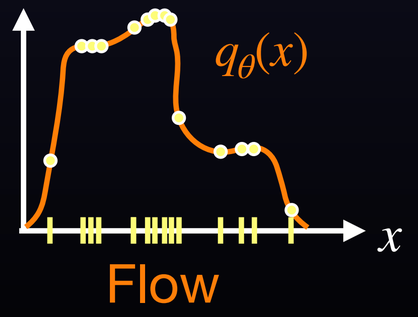
Forward KL Divergence (fKLD)

$$L_{\text{fKLD}} = D_{\text{KL}}(p \,\|\, q_\theta) = \mathbb{E}_{x\sim p(x)}\left[\log \frac{p(x)}{q_\theta(x)}\right] = -\sum_{i=1}^{N} \log q_\theta(x_i) + \text{const.}$$





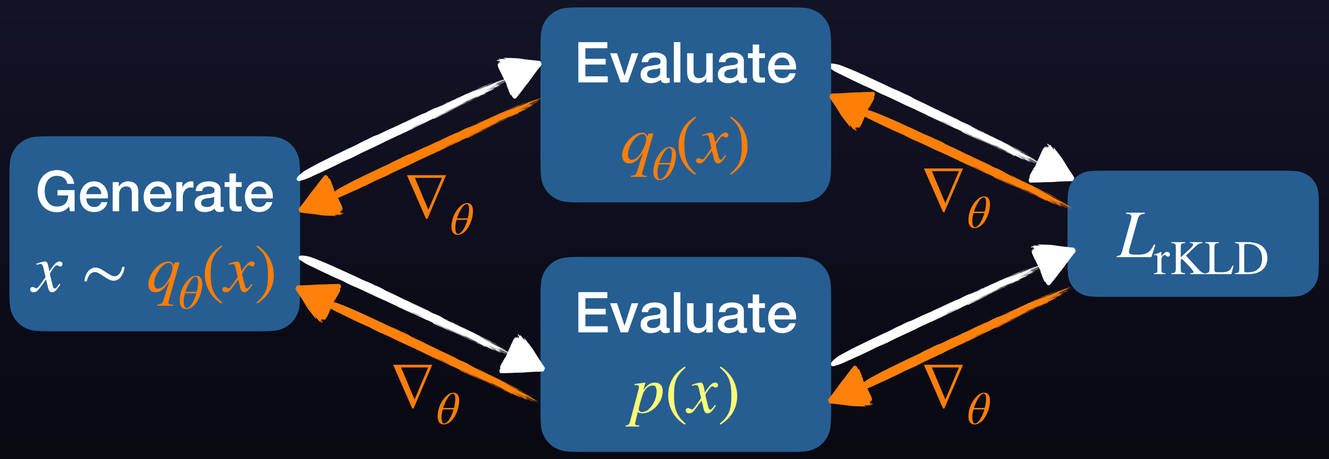
→ Expensive to generate training data

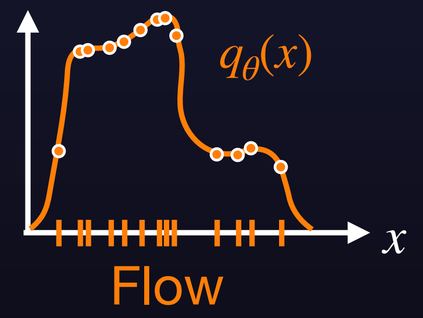


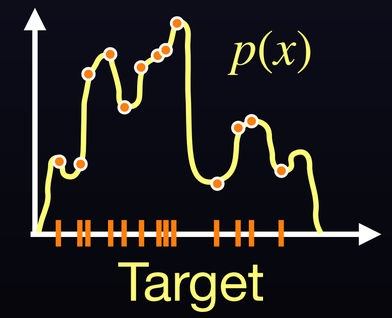
Training with Density Evaluation

Reverse KL Divergence (rKLD)

$$L_{\text{rKLD}} = D_{\text{KL}}(q_\theta \,\|\, p) = \mathbb{E}_{x\sim q_\theta(x)}\left[\log \frac{q_\theta(x)}{p(x)}\right]$$





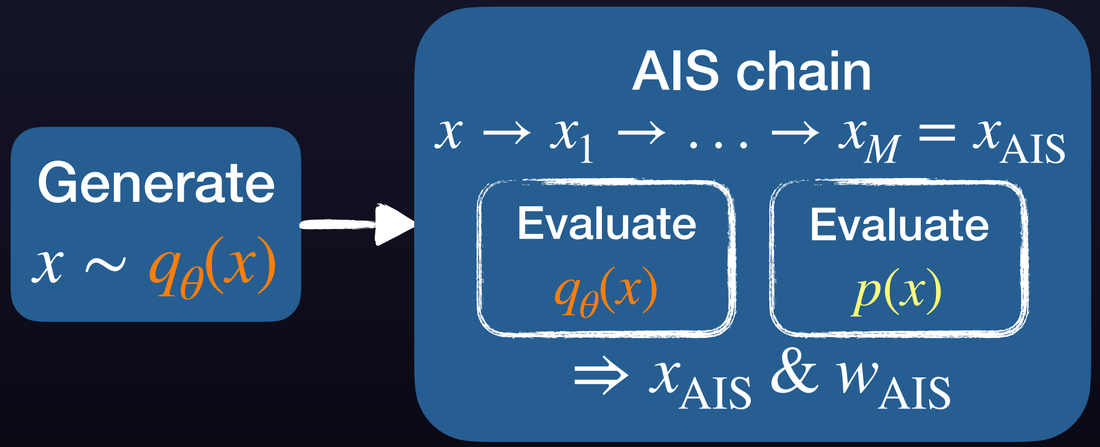


→ Requires differentiable target distribution $p(x)$

Training with Density Evaluation

Flow Annealed Importance Sampling Bootstrap (FAB)

Improve flow samples $x \sim q_\theta(x)$ with Annealed Importance Sampling (AIS)



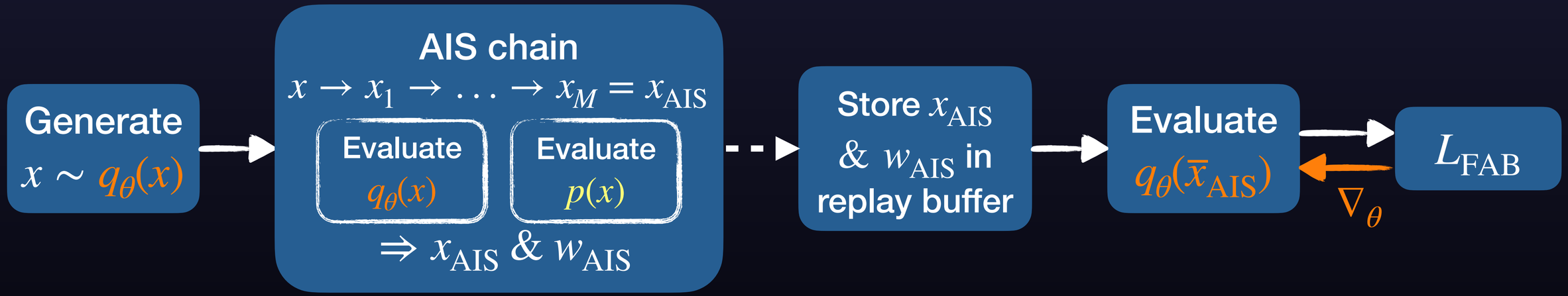
AIS chain implemented via Hamiltonian Monte Carlo (HMC)
→ Requires differentiable target $p(x)$

Midgley, Stimper, et al., “Flow Annealed Importance Sampling Bootstrap”, ICLR, 2023.

Training with Density Evaluation

Flow Annealed Importance Sampling Bootstrap (FAB)

Improve flow samples $x \sim q_\theta(x)$ with Annealed Importance Sampling (AIS)



$$L_{\text{FAB}} = -\sum_{i=1}^{N} \frac{\bar{w}_{\text{AIS}}^{(i)}}{\sum_j \bar{w}_{\text{AIS}}^{(j)}} q_\theta\left(\bar{x}_{\text{AIS}}^{(i)}\right)$$

→ minimizes variance of importance weights

Midgley, Stimper, et al., “Flow Annealed Importance Sampling Bootstrap”, ICLR, 2023.

Differentiable target distributions

Examples

Recent work in particle physics: differentiable implementations of $p(x)$

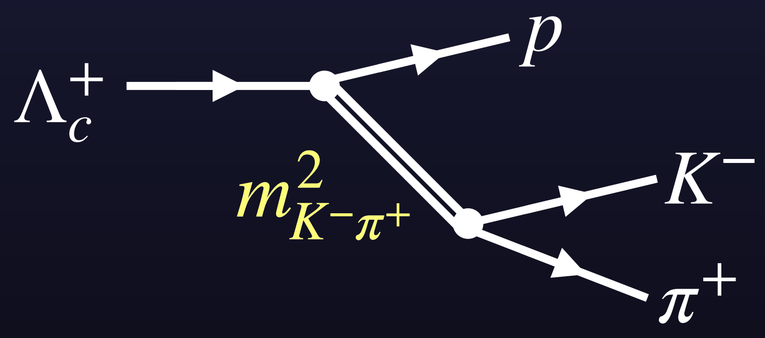
- **ComPWA:** $\Lambda_c^+ \to pK^-\pi^+$ (2D)
- **MadJAX:** $e^+e^- \to t\bar{t}, t\bar{t} \to W^+b, \bar{t} \to W^-\bar{b}$ (8D)

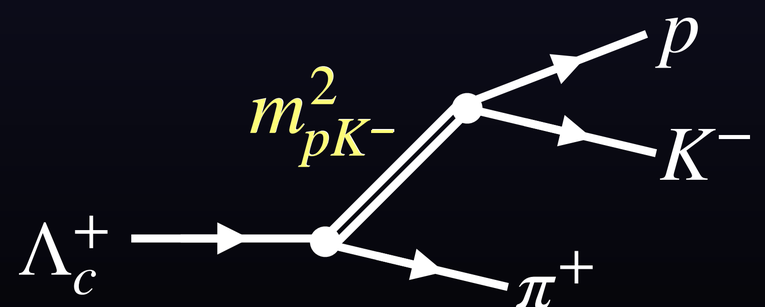
Common Partial Wave Analysis: A collaboration-independent organisation for amplitude analysis software, doi 10.5281/zenodo.6908150

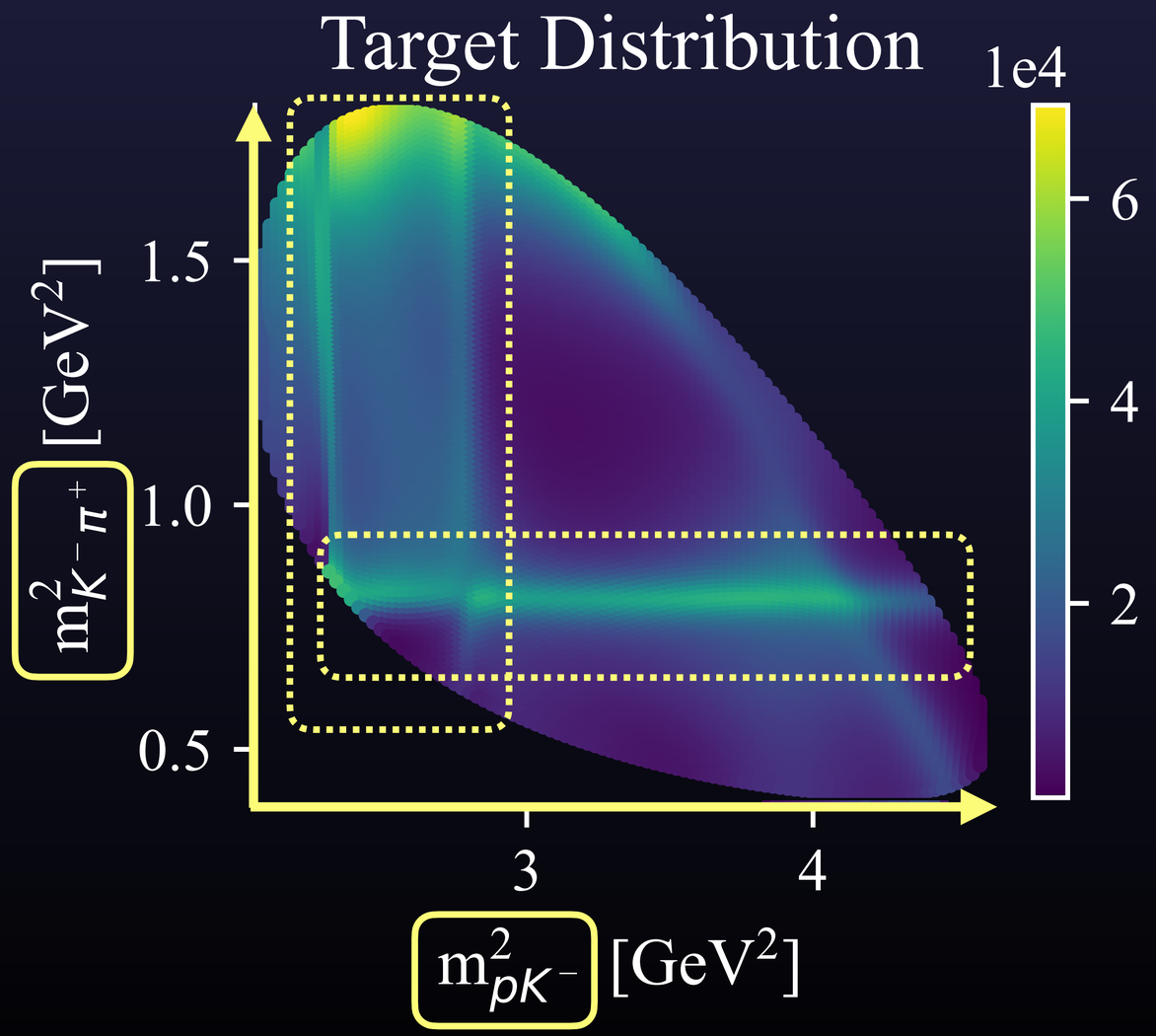
Heinrich and Kagan, “Differentiable Matrix Elements with MadJax.” J. Phys. Conf., 2023.

2D: $\Lambda_c^+ \to pK^-\pi^+$

What do we see?



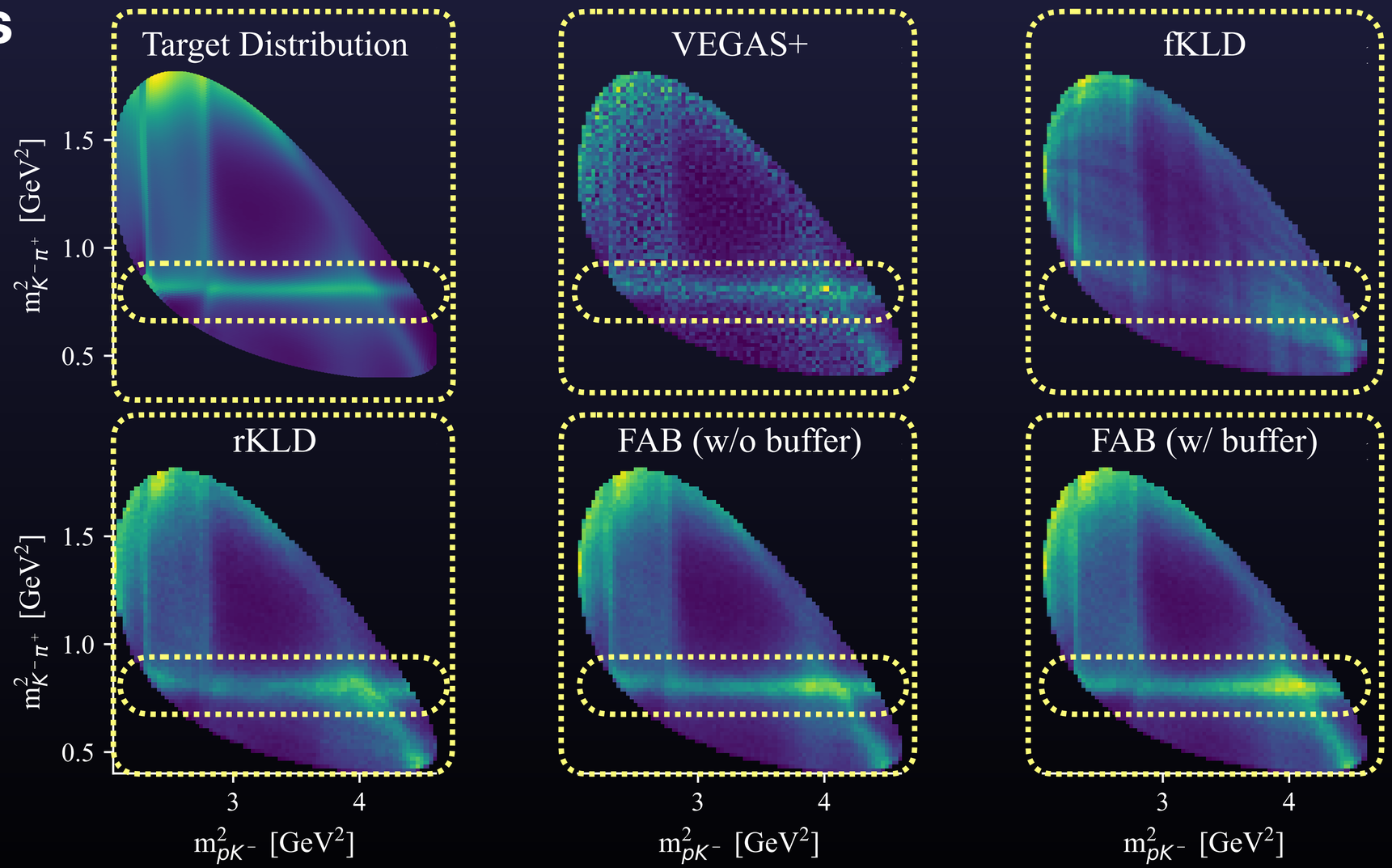




LHCb, “Amplitude analysis of the $\Lambda_c^+ \to pK^-\pi^+$ decay and Λ_c^+ baryon polarization measurement in semileptonic beauty hadron decays.”, Phys. Rev. D 108, 2023.

2D: $\Lambda_c^+ \to pK^-\pi^+$

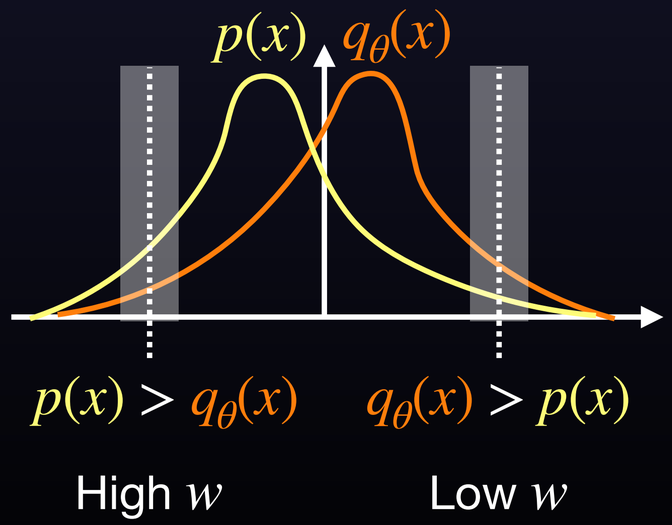
Histograms



How to compare?

- Efficiency ϵ

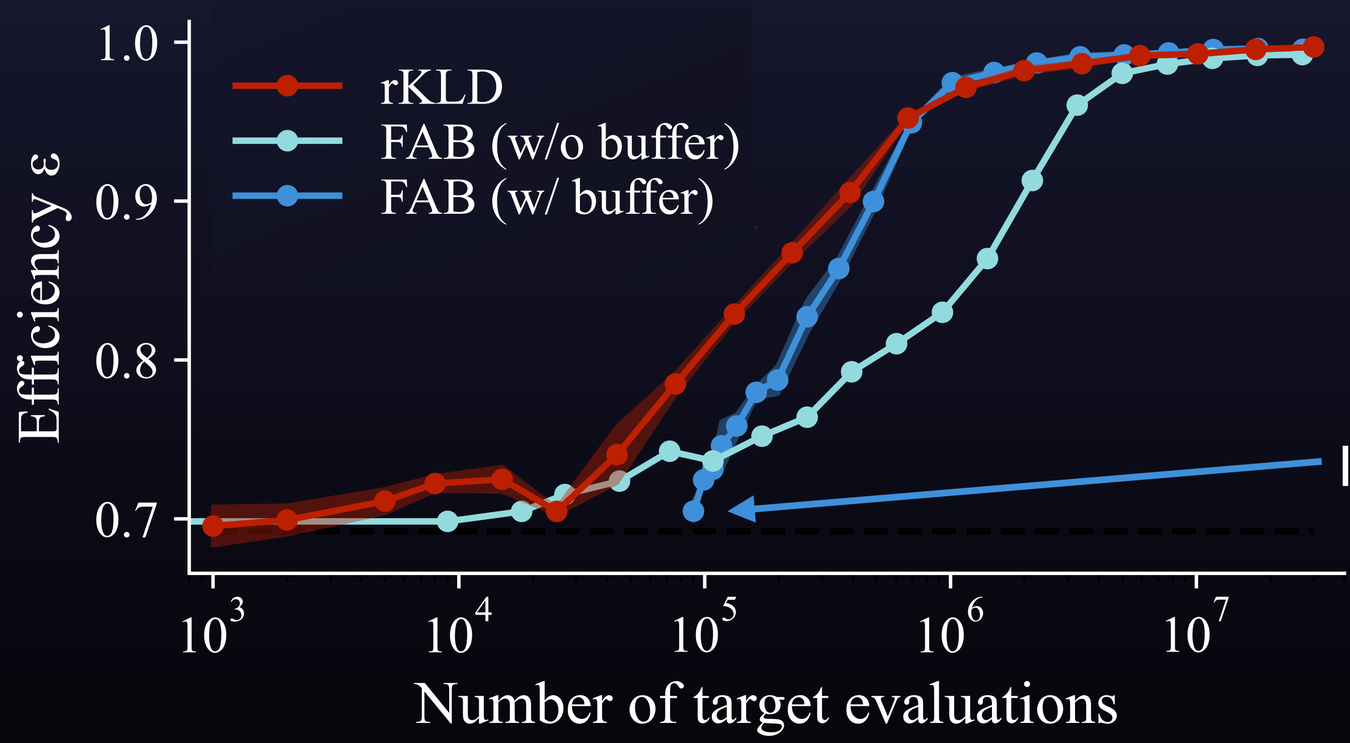
$$\epsilon = \frac{1}{N}\frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2} \in [0,1] \quad \text{with importance weight} \quad w_i = \frac{p(x_i)}{q_\theta(x_i)}$$



$$p(x) \neq q_\theta(x) \rightarrow \epsilon \ll 1$$

2D: $\Lambda_c^+ \to pK^-\pi^+$

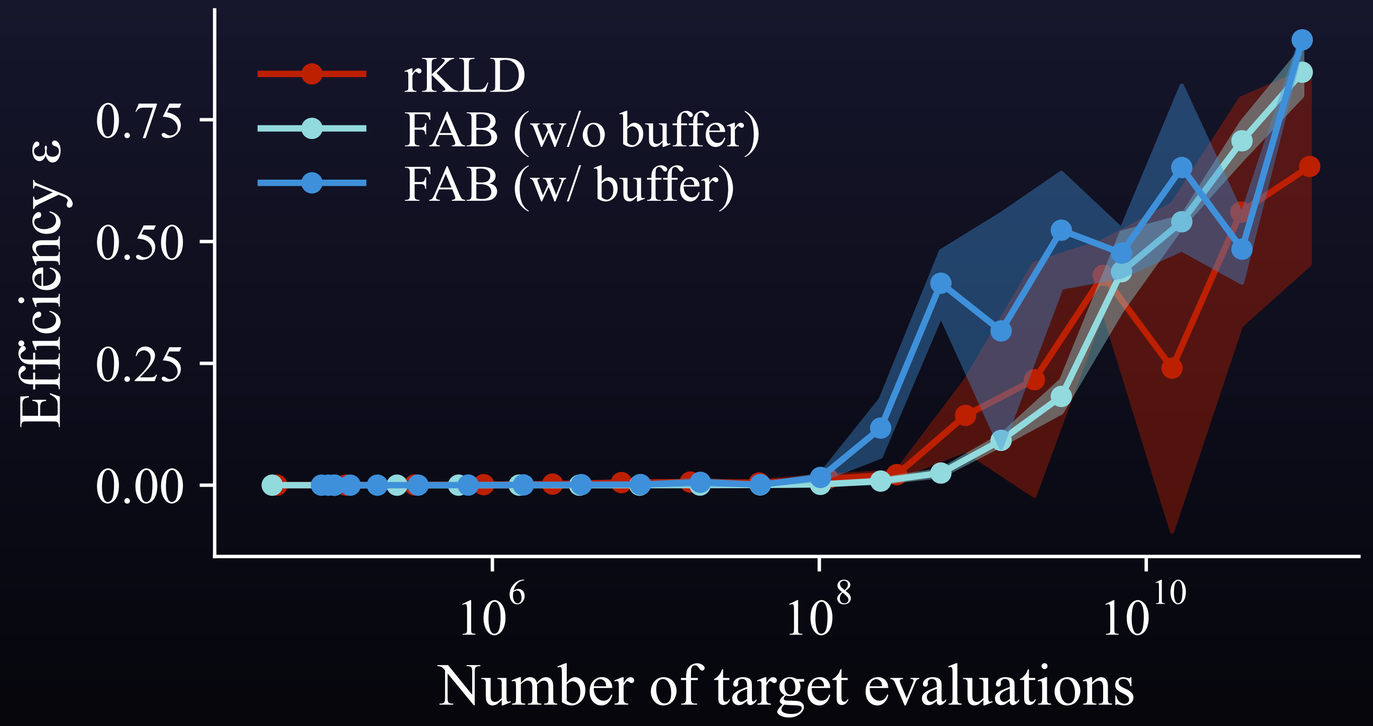
Efficiency vs. Number of target evaluations



Initial filling of replay buffer

8D: $e^+e^- \to t\bar{t}, t\bar{t} \to W^+b, \bar{t} \to W^-\bar{b}$

Efficiency vs. Number of target evaluations



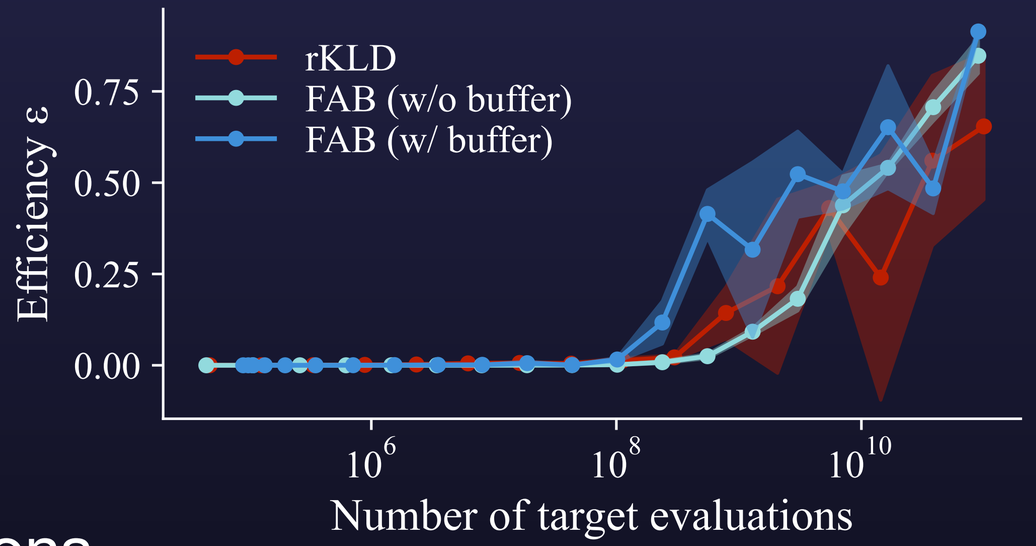
Take-Aways

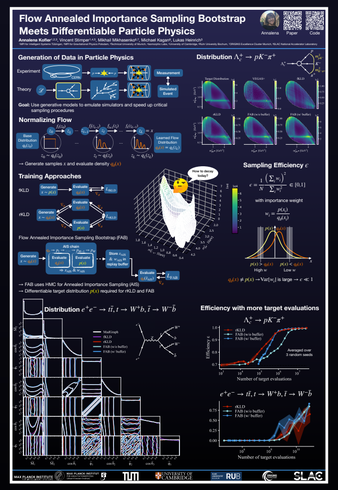
FAB (w/ buffer) …

… successfully adapted to particle physics

… outperforms other methods in high dimensions

… achieves higher sampling efficiency with fewer target evaluations





Visit my poster!
(16:15 - 17:25)

Paper

Code

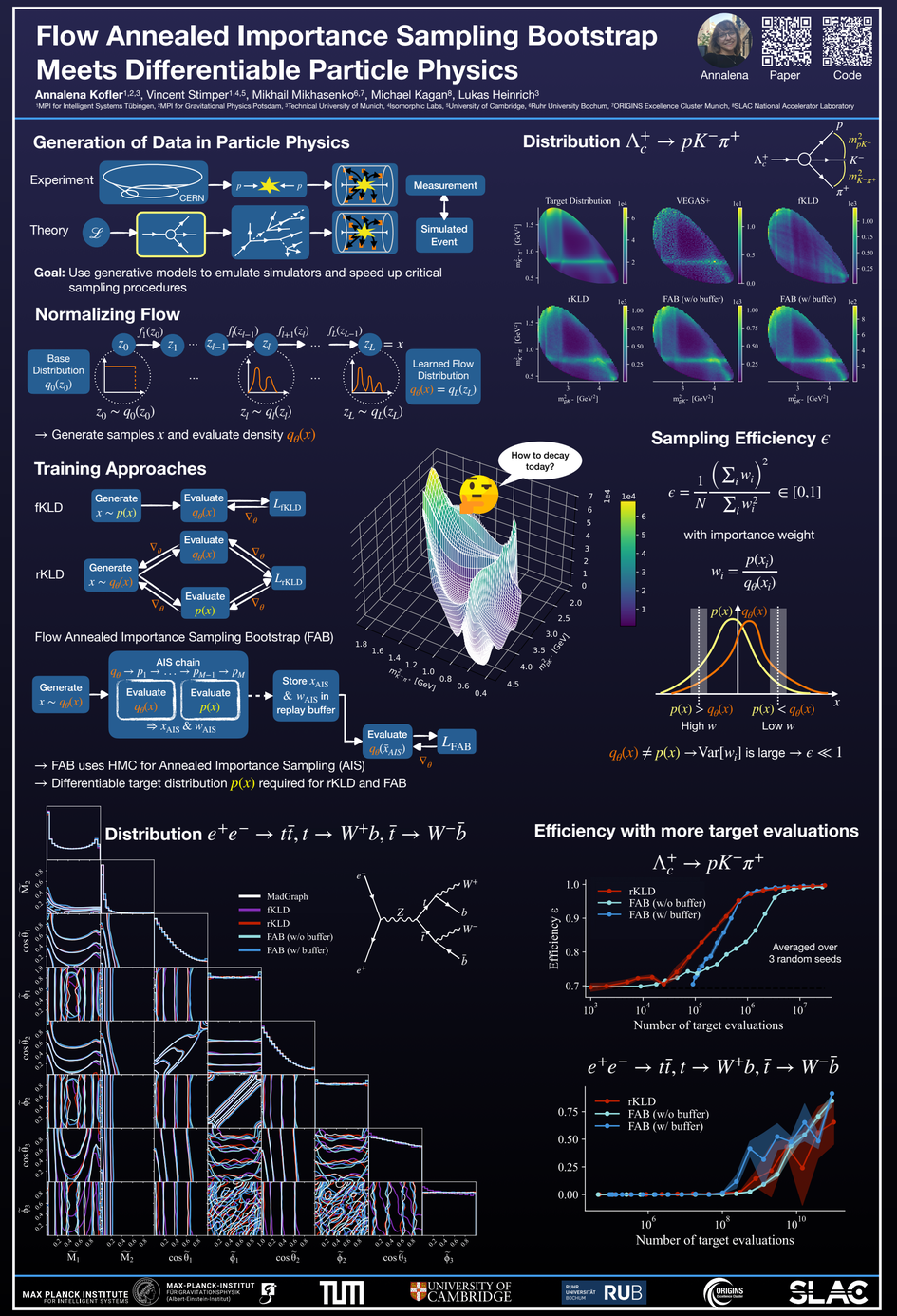
Thanks to WiML for sponsoring my travel!

Do you have any questions?

You want to have an in-depth discussion?
→ Visit my poster (16:15 - 17:25)

Paper

Code



References

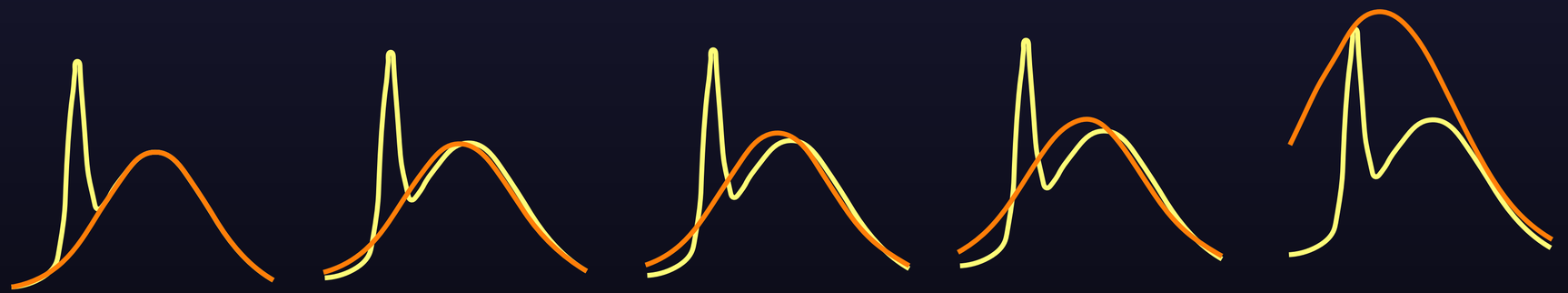
- ATLAS Software and Computing HL-LHC Roadmap, 2022.
- Rezende and Mohamed, “Variational Inference with Normalizing Flows.” ICML’15.
- Durkan et al., “Neural Spline Flows.”, NeurIPS’19.
- LHCb, “Amplitude analysis of the $\Lambda_c^+ \to pK^-\pi^+$ decay and Λ_c^+ baryon polarization measurement in semileptonic beauty hadron decays.”, Phys. Rev. D 108, 2023.
- Heinrich and Kagan, “Differentiable Matrix Elements with MadJax.” J. Phys. Conf., 2023.
- Midgley, Stimper, et al., “Flow Annealed Importance Sampling Bootstrap”, ICLR, 2023.

Appendix

Flow Annealed Importance Sampling Bootstrap (FAB)

What is special about the loss function?

- α-divergence $\quad D_\alpha(p\|q_\theta) = -\frac{1}{\alpha(1-\alpha)}\int p(x)^\alpha q_\theta(x)^{1-\alpha}\,\mathrm{d}x$



$\alpha = -\infty$ $\quad$ $\alpha \to 0$ rKLD $\quad$ $\alpha = 0.5$ $\quad$ $\alpha \to 1$ fKLD $\quad$ $\alpha = \infty$

$$D_{\alpha=2}(p\|q_\theta) \propto \int \frac{p(x)^2}{q_\theta(x)}\,\mathrm{d}x \quad \xrightarrow{\text{Approximate}} \quad L_{\text{FAB}} = -\sum_{i=1}^{N} \frac{\bar{w}_{\text{AIS}}^{(i)}}{\sum_j \bar{w}_{\text{AIS}}^{(j)}} q_\theta\left(\bar{x}_{\text{AIS}}^{(i)}\right)$$

Flow Annealed Importance Sampling Bootstrap (FAB)

How does Annealed Importance Sampling (AIS) work?

- Interpolate between two distributions p_0 and p_M
  - Intermediate distribution $\log p_m(x) = \beta_m \log p_0(x) + (1 - \beta_m) \log p_M(x)$

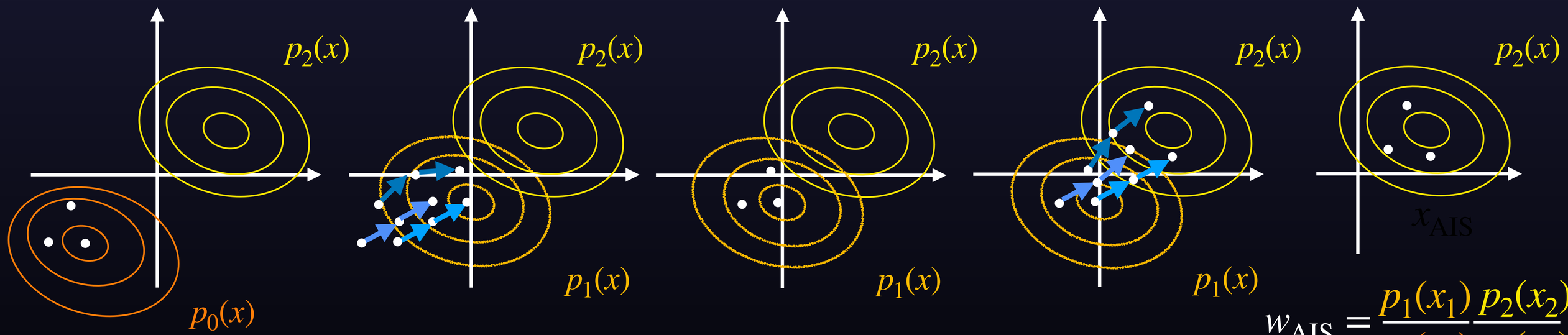
$$\beta_0 = 1 > \ldots > \beta_m > \ldots > \beta_M = 0$$

- Move samples from previous p_m to next p_{m+1} via HMC
  - Track importance weight

$$w_{\text{AIS}} = \frac{p_1(x_1)}{p_0(x_1)} \frac{p_2(x_1)}{p_1(x_1)} \cdots \frac{p_M(x_M)}{p_{M-1}(x_M)}$$

Flow Annealed Importance Sampling Bootstrap (FAB)

How does Annealed Importance Sampling (AIS) work?



$$w_{\text{AIS}} = \frac{p_1(x_1)}{p_0(x_1)} \frac{p_2(x_2)}{p_1(x_2)}$$

Midgley, Stimper, et al., “Flow Annealed Importance Sampling Bootstrap”, ICLR, 2023.

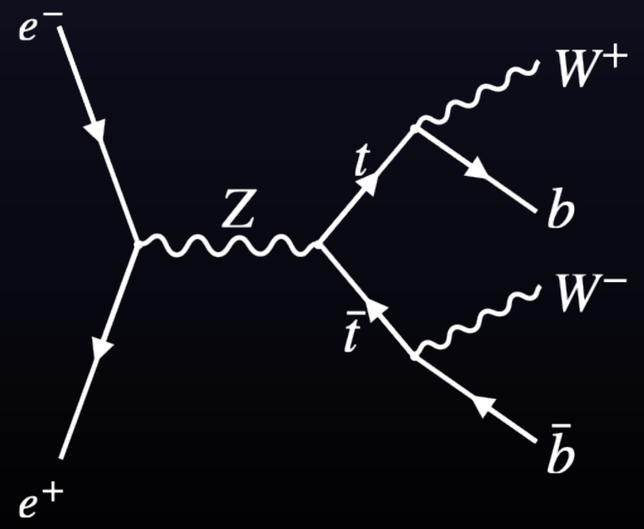
Flow Annealed Importance Sampling Bootstrap (FAB)

Replay buffer

Algorithm 7: Prioritized Replay Buffer for FAB

Initialize flow $q_\theta(x)$
Initialize replay buffer to a fixed maximum size
for $i = 0$ *to* K **do**
  # Generate batch B of AIS samples and add to buffer:
  Sample batch $x^{(1:B)}$ from flow $q_\theta(x)$ and evaluate $\log q_\theta(x^{(1:B)})$
  Run AIS with target d^2/q_θ starting from $x^{(1:B)}$ and $\log q_\theta(x^{(1:B)})$ to get $x_{\text{AIS}}^{(1:B)}$ and $\log w_{\text{AIS}}^{(1:B)}$
  Add $x_{\text{AIS}}^{(1:B)}$, $\log w_{\text{AIS}}^{(1:B)}$, and $\log q_\theta(x_{\text{AIS}}^{(1:B)})$ to replay buffer
  for $j = 1$ *to* J **do**
    # Sample batch b from buffer and update flow:
    Sample $x_{\text{AIS}}^{(1:b)}$ and $\log q_\theta(x_{\text{AIS}}^{(1:b)})$ from buffer with probability $p \propto w_{\text{AIS}}^{(1:b)}$
    Calculate $\log w_{\text{correction}}^{(1:b)} = \log q_{\theta_{\text{old}}}(x_{\text{AIS}}^{(1:b)}) + \text{stopgrad}(\log q_\theta(x_{\text{AIS}}^{(1:b)}))$
    Update $\log w_{\text{AIS}}^{(1:b)} \leftarrow \log w_{\text{AIS}}^{(1:b)} + \log w_{\text{correction}}^{(1:b)}$ in buffer
    Update $\log q_{\theta_{\text{old}}}(x_{\text{AIS}}^{(1:b)}) \leftarrow \log q_\theta(x_{\text{AIS}}^{(1:b)})$
    Calculate loss $S'(\theta) = -\frac{1}{N}\sum_i w_{\text{correction}}^{(i)} \log q_\theta(x_{\text{AIS}})^{(1:b)}$
    Perform gradient descent on $S'(\theta)$ to update θ

Results: $e^+e^- \to t\bar{t}$, $t \to W^+b, \bar{t} \to W^-\bar{b}$ (8D)



8D: $e^+e^- \to t\bar{t}$,
$t\bar{t} \to W^+b, \bar{t} \to W^-\bar{b}$

Corner Plot