

# Neural (Quasi-)Probabilistic Likelihood Ratio Estimation

Matthew Drnevich & Stephen Jiggins

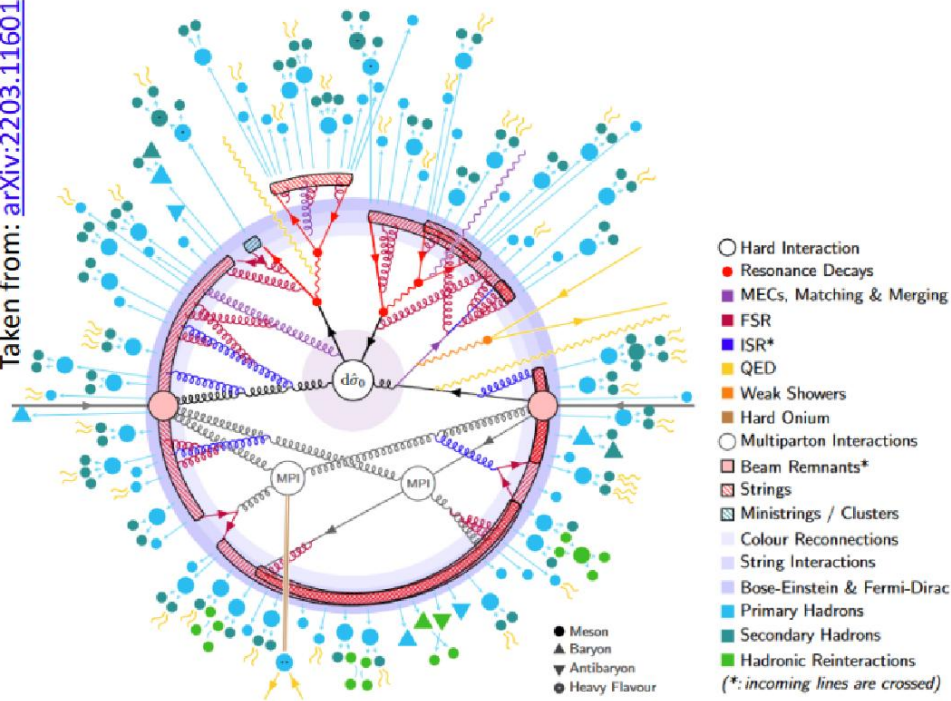


NYU



# Particle Physics ↔ ML

Taken from: [arXiv:2203.11601](https://arxiv.org/abs/2203.11601)



**Monte Carlo/Statisticians Paradigm:**  
Probability Density Function

$$I := \int_a^b p(\vec{x}) d\vec{x}$$

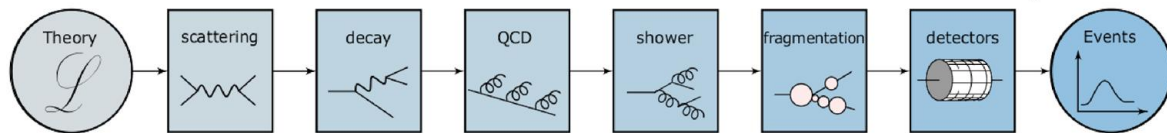
→ Sampling of a *probability density function (pdf)* called  $\mathbf{p}(\vec{x})$

**Particle Physicists Theoretical Paradigm:**  
Quantum Field Theory

$$\hat{\sigma}_{a,b \rightarrow m} = \int_{\Omega} d\hat{\sigma}_{a,b \rightarrow m} = \int_{\Omega} |\mathcal{M}_{a,b \rightarrow m}|^2 d\Phi_m(\vec{x})$$

**Where:**

$$d\Phi_m(\vec{x}) = \prod_{i=1}^m \frac{d^3 \vec{p}_i}{(2\pi^3)2E_i} \delta^4 \left( p_a + p_b - \sum_{j=1}^m p_j \right)$$



# Simulation-Based Inference

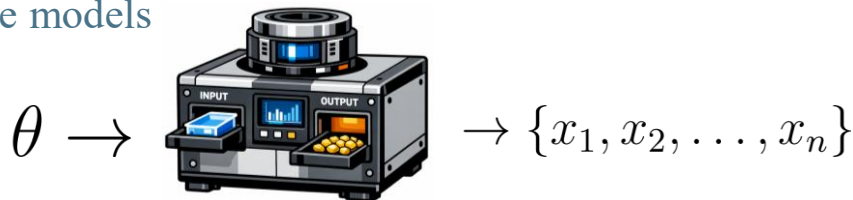
We have data sampled from a physics model determined by some parameters

$$x \sim |\mathcal{M}(x|\theta)|^2$$

and we want to infer the physical parameters that our data describes implicitly

$$\{x_1, x_2, \dots, x_n\} \implies \theta = ?$$

with only access to simulation/generative models



How can we construct meaningful statistical models and perform hypothesis testing?

$$\begin{aligned} H_0 : \theta \in \Theta_0, & \quad \text{reject } H_0? \\ H_1 : \theta \in \Theta_1, & \end{aligned}$$

**The likelihood-ratio test is the most powerful (Neyman-Pearson lemma)**

# Likelihood Ratio Tests

Likelihood ratio tests are state-of-the-art and currently employed by analyses

$$\lambda_{LR} = -2 \ln \left[ \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)} \right]$$

But can we do this in a continuous, multivariate way?

- We can try to model the densities:  $p(x|\theta)$
- **Or we can try to model the ratio:**  $r(x|\theta_0, \theta_1) = \frac{p(x|\theta_0)}{p(x|\theta_1)}$

**This talk!**

The reference value can be somewhat arbitrary since

$$\ln \left[ \frac{p(x|\theta_0)}{p(x|\hat{\theta})} \right] = \ln \left[ \frac{p(x|\theta_0)p(x|\theta_1)}{p(x|\hat{\theta})p(x|\theta_1)} \right] = \ln r(x|\theta_0, \theta_1) - \ln r(x|\hat{\theta}, \theta_1)$$

**So how do we estimate the ratio?**

# Neural Likelihood Ratio Estimation

# General Setup

1. Start with a neural network

$$s(x; \theta)$$

2. Choose some loss functional

$$L[s(x; \theta_t)]$$

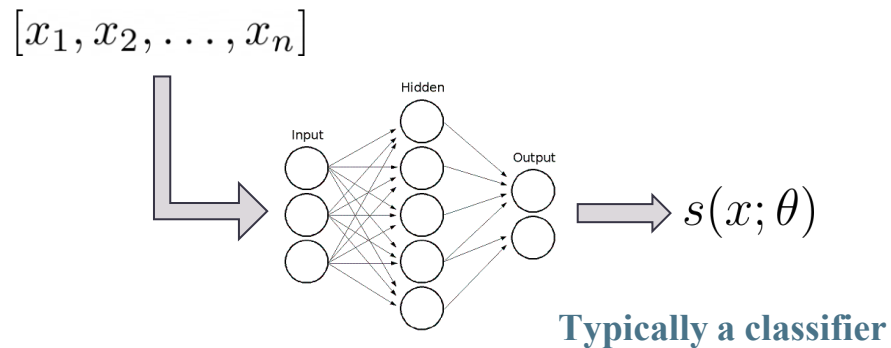
3. Find the minimum using gradient descent

$$\theta_{t+1} = \theta_t - \eta_t \nabla_{\theta} L[s(x; \theta_t)]$$

4. The optimal function is related to the density ratio through some transformation

$$s(x; \theta^*) = \operatorname{argmin}_{\theta} L[s(x; \theta)] \implies T(s(x; \theta^*)) = r(x)$$

subject to certain conditions (see [\[arxiv:2512.19913\]](#))



# Loss Functions

General Loss Function

$$L[s] = -\int dx [p_0 A(s) + p_1 B(s)]$$

Loss functions have a general form:

$$L[s] = -\int dx [p(x|\theta_0)A(s(x)) + p(x|\theta_1)B(s(x))]$$

Optimal functions satisfy the “Ratio Trick” equation:

$$\frac{\delta L}{\delta s} = 0 \iff T(s^*(x)) := -\frac{A'(s^*(x))}{B'(s^*(x))} = \frac{p(x|\theta_1)}{p(x|\theta_0)} = r(x|\theta_0, \theta_1)$$

While many loss functions exist, including ones that directly estimate the ratio, we’ll focus on classifier-based ones where  $s(x) \in (0, 1)$

# Classification Loss

If we have access to generative models, but not the density functions, then we can do classification

Use Bayes' Law to re-write  $p(x|\theta)$

$$p(x|\theta) = \frac{p(\theta|x)p(x)}{p(\theta)}$$

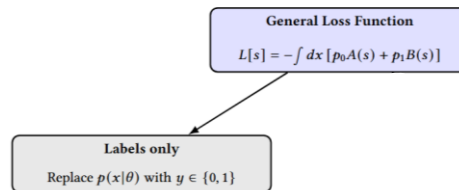
Sample the same number from each  $p(x|\theta_0)$  and  $p(x|\theta_1)$  so  $p(\theta) = 1/2$

$$p(\theta_0|x) \mapsto \mathbb{E}_{y \sim p(y|x)} [1 - y], \quad p(\theta_1|x) \mapsto \mathbb{E}_{y \sim p(y|x)} [y]$$

Then you get a binary classification loss

$$L[s] = -\mathbb{E}_{(x,y) \sim p(x,y)} [(1 - y)A(s(x)) + yB(s(x))]$$

Replaced the integral over unknown densities with an expectation over data



# CARL Models

Now choose the functions A, B

$$A(s) = \ln(1 - s), \quad B(s) = \ln(s)$$

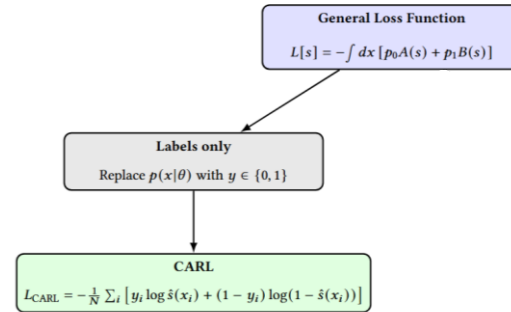
Then you get Binary Cross Entropy

$$L_{BCE}[s] = -\mathbb{E}_{(x,y) \sim p(x,y)} [(1 - y) \ln(1 - s(x)) + y \ln(s(x))]$$

With the “Ratio Trick”:  $r(x|\theta_0, \theta_1) = \frac{s^*(x)}{1 - s^*(x)}$

CARL combines this loss function with calibration methods to produce meaningful probabilities

This is state-of-the-art without additional information



# “Mining Gold”

General Loss Function  
 $L[s] = -\int dx [p_0 A(s) + p_1 B(s)]$

Joint simulator info  
Access to  $r(x, z|\theta)$  and  $\nabla_\theta \log r(x, z|\theta)$

What if we can extract additional information from the simulator?

Often we have access to the probabilities/ratios defined on some latent space, e.g. parton-level  $p(z|\theta)$

These can be used to compute auxiliary quantities

**The joint ratio**  $r(x, z|\theta_0, \theta_1) = \frac{p(x, z|\theta_1)}{p(x, z|\theta_0)}$

**The joint score**  $t(x, z|\theta_0) = \nabla_\theta \log p(x, z|\theta)|_{\theta_0}$

Computed during simulation



This information can be used to improve the loss functions

[\[1805.12244\] Mining gold from implicit models to improve likelihood-free inference](#)

# ALICE Loss

Return to the classification loss function, but don't use class labels – use the joint ratio instead

$$p(\theta_0|x) \mapsto \frac{1}{r(x, z|\theta_0, \theta_1) + 1}, \quad p(\theta_1|x) \mapsto \frac{r(x, z|\theta_0, \theta_1)}{r(x, z|\theta_0, \theta_1) + 1}$$

Then replace the expectation over  $y$  with  $z$

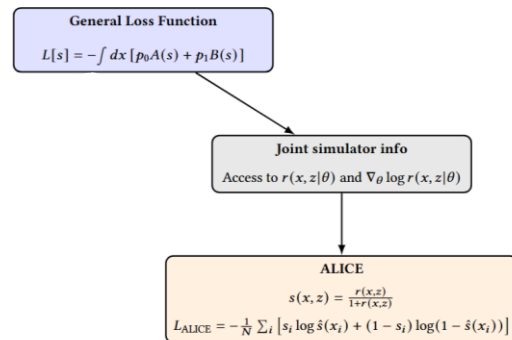
$$L_{ALICE}[s] = -\mathbb{E}_{(x,z) \sim p(x,z)} [(1 - s(x, z|\theta_0, \theta_1)) \ln(1 - s(x)) + s(x, z|\theta_0, \theta_1) \ln(s(x))]$$

where

$$s(x, z|\theta_0, \theta_1) = p(\theta_1|x, z) = \frac{r(x, z|\theta_0, \theta_1)}{r(x, z|\theta_0, \theta_1) + 1}$$

**This leads to improved performance due to reduced variance**

[\[1808.00973\] Likelihood-free inference with an improved cross-entropy estimator](#)



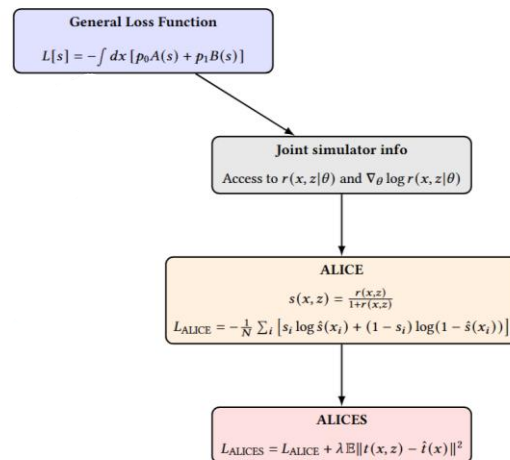
# ALICES Loss

Take the ALICE loss and add regression on the score, which adds local differential behavior to the learning

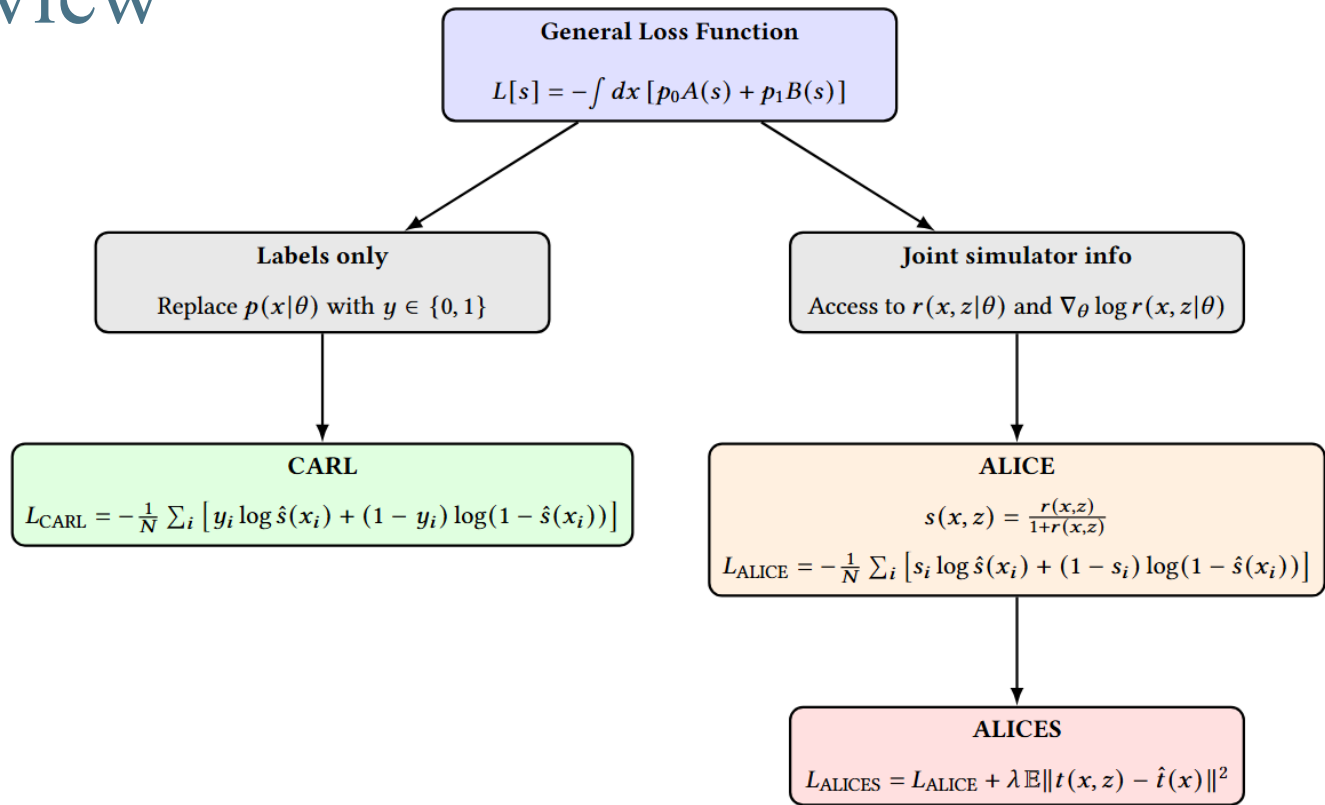
$$t(x, z | \theta_0) = \nabla_{\theta} \log p(x, z | \theta) \Big|_{\theta_0}$$

This can further improve the sample efficiency of learning

$$L_{ALICES} = L_{ALICE} + \alpha(1 - y) \left| t(x, z | \theta_0, \theta_1) - \nabla_{\theta} \log \left( \frac{1 - \hat{s}(x | \theta, \theta_1)}{\hat{s}(x | \theta, \theta_1)} \right) \Big|_{\theta_0} \right|^2$$



# Overview



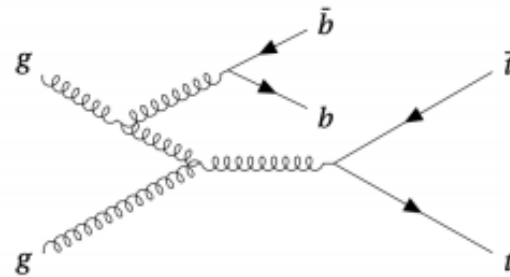
\*There are many more methods not mentioned here, but see the linked papers on previous slides

# Negative Weights

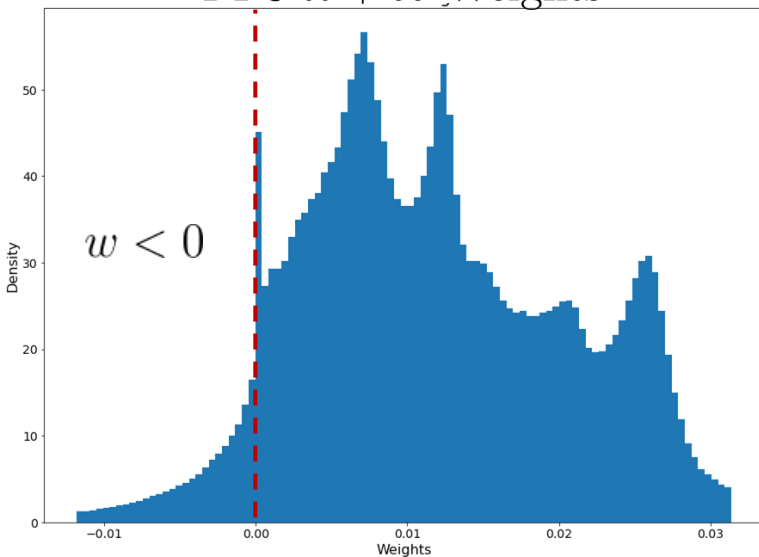
# Negative Weights

In practice, models trained on negatively-weighted data can struggle with convergence

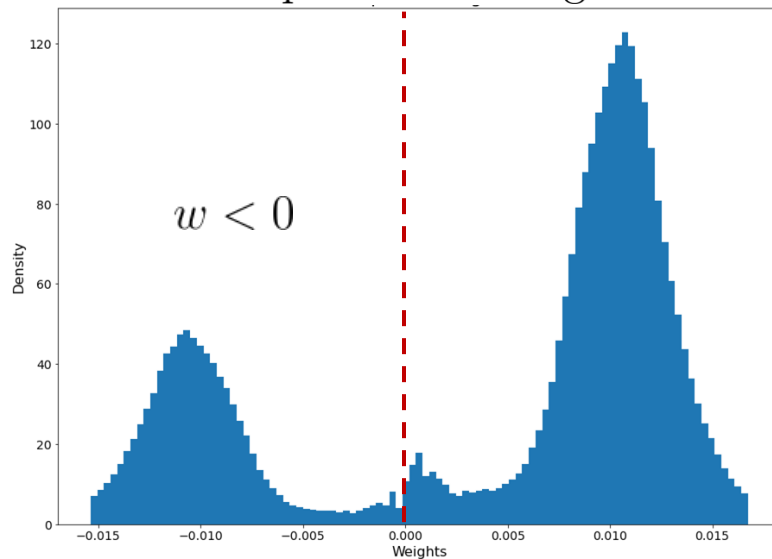
Some physics samples can have a large proportion of negatively weighted events!



PP8  $t\bar{t} + b\bar{b}$  Weights



Sherpa  $t\bar{t} + b\bar{b}$  Weights



# Where's the Problem?

Updating model parameters via gradient descent

$$\theta^{t+1} = \theta^t - \frac{\eta}{N} \sum_{i=1}^N \nabla_{\theta^t} L(x_i, y_i; s_{\theta^t}) w_i$$

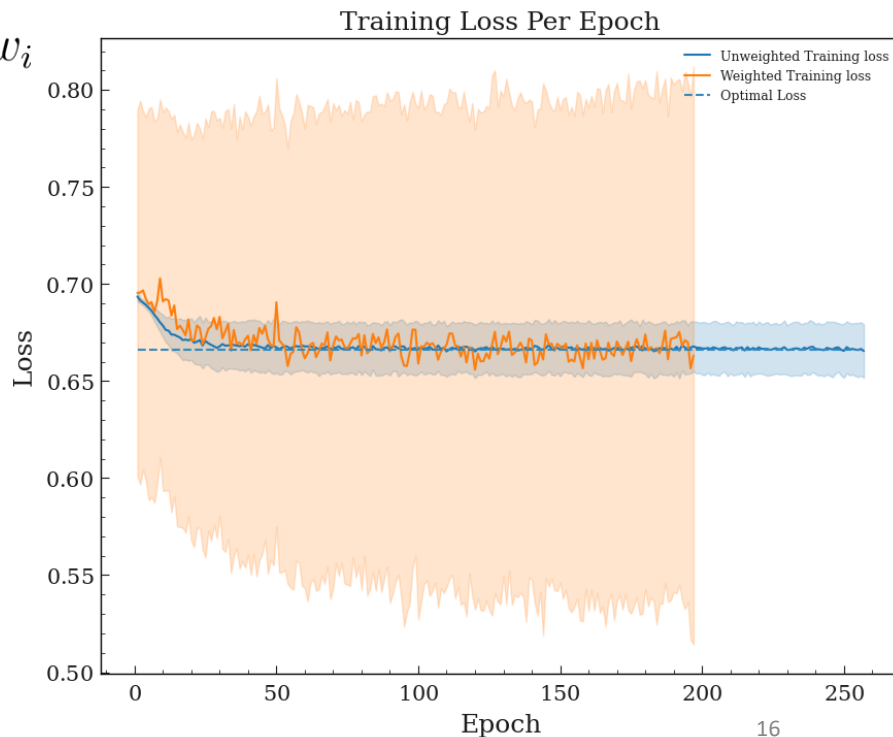
Training on unweighted data

$$\text{Var}(\theta^{t+1} | X, Y) = 0$$

Training on weighted data

$$\text{Var}(\theta^{t+1} | X, Y) \propto \text{Var}(W | X, Y)$$

**This is independent  
of the loss function**



# Overview of the Problem

Monte Carlo data often includes negatively weighted events

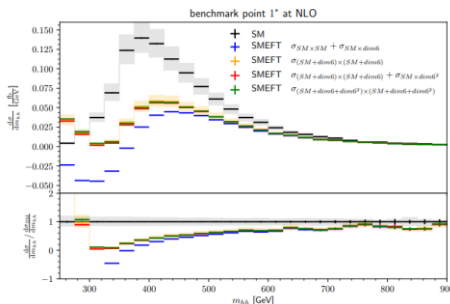
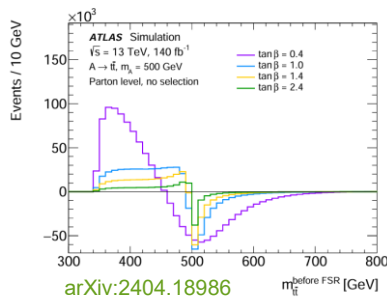


← **Explicit BSM QFT:**  
Exact theoretical extension of the SM with new BSM physics.  
E.g. 2HDMs or MSSM

← **Effective Field Theory:**  
Effective interaction vertices of unknown form can extend the SM (SMEFT) that introduces physics at some new scale  $\Lambda$ :

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{C_i^{(6)}}{\Lambda^2} \mathcal{O}_i^{\text{dim}6} + \mathcal{O}\left(\frac{1}{\Lambda^3}\right)$$

← **Monte Carlo estimates of SM  $\sigma$ 's:**  
Merging/matching of  $\mathcal{M}^n$  and  $\mathcal{M}^{n+1}$  matrix elements from MC estimates of p-p collision; e.g. MC@NLO subtraction method



5

$$\mathcal{M}^n \leftrightarrow \mathcal{M}^{n+1}$$

1. Theoretical extensions of the Standard Model (SM) that interfere with SM physics

$$\mu S + \sqrt{\mu} I + B = (\mu - \sqrt{\mu}) S + \sqrt{\mu} (S + I) + B$$

2. Effects from truncating perturbative series at a fixed order of the new physics scale  $\Lambda$

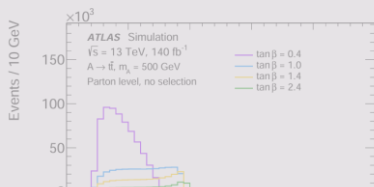
3. Numerical corrections to sampling imperfections

# Overview of the Problem

Monte Carlo data often includes negatively weighted events



← **Explicit BSM QFT:**  
Exact theoretical extension of the SM with new BSM physics.  
E.g. 2HDMs or MSSM

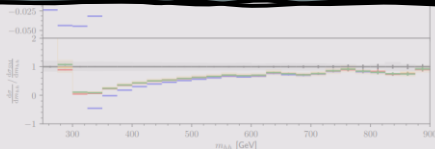


1. Theoretical extensions of the Standard Model (SM) that interfere with SM physics

**Can we develop new machine learning techniques to safely train on negatively weighted samples?**

← **Effective unknown SM physics at some new scale  $\Lambda$ :**

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{C_i^{(6)}}{\Lambda^2} \mathcal{O}_i^{\text{dim6}} + \mathcal{O}\left(\frac{1}{\Lambda^3}\right)$$



[arXiv:2204.1304](https://arxiv.org/abs/2204.1304)  
5

← **Monte Carlo estimates of SM  $\sigma$ 's:**  
Merging/matching of  $\mathcal{M}^n$  and  $\mathcal{M}^{n+1}$  matrix elements from MC estimates of p-p collision; e.g. MC@NLO subtraction method

$$\mathcal{M}^n \leftrightarrow \mathcal{M}^{n+1}$$

3. Numerical corrections to sampling imperfections

$(S + I) + B$

relative new

physics scale  $\Lambda$

# Neural Quasiprobabilistic Density Ratio Estimation

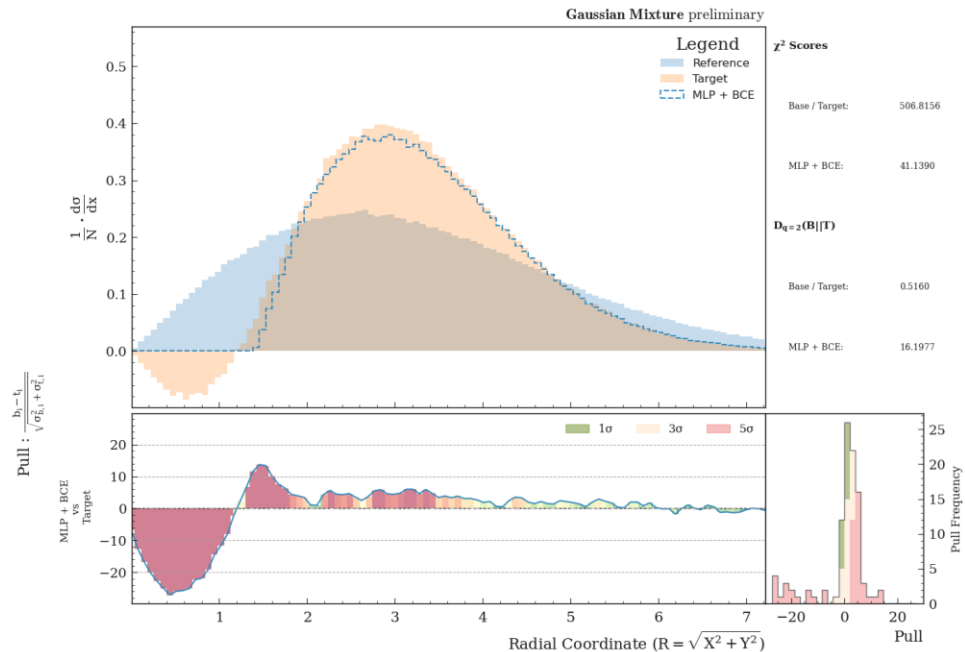
# Classical Limitations

Traditional methods are restricted to nonnegative values

$$\hat{r}(x) = \frac{s(x)}{1 - s(x)} \geq 0$$

Current approaches will fail whenever negative density regions of phase space are present

## Toy Model Demonstration



# Solution 1: Extending CARL

# A New Loss Function: REVERT

We introduce a new loss function

$$\mathcal{L}(s(x), y) = ys - (1 - y)(\log s(x) + \log(1 - s(x)))$$

and train a binary classifier

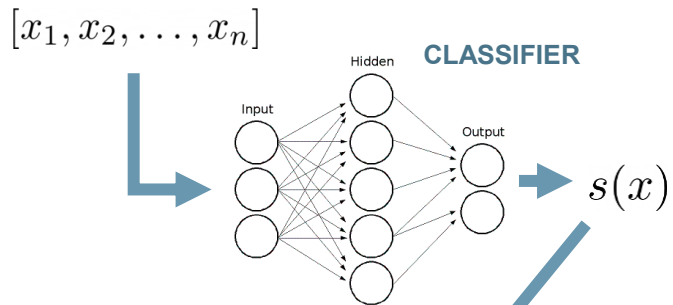
$$s(x) \in (0, 1)$$

With a new “ratio trick”

$$\hat{r}(x) = \frac{1}{s(x)} + \frac{1}{s(x) - 1}$$

which is capable of representing any density ratio

$$r(x) \in (-\infty, \infty)$$



$$\hat{r}(x) = \frac{1}{s(x)} + \frac{1}{s(x) - 1} \approx \frac{q(x|\theta_1)}{q(x|\theta_0)}$$

# Solution 2: Ratio of Signed Mixtures Model

# Breaking up the Learning Task

Decompose the quasi-pdfs into signed mixture models

$$q(x|\theta) = c(\theta)p(x|\theta, w \geq 0) - (c(\theta) - 1)p(x|\theta, w < 0)$$

# Breaking up the Learning Task

Decompose the quasi-pdfs into signed mixture models

$$q(x|\theta) = c(\theta) p(x|\theta, w \geq 0) - (c(\theta) - 1) p(x|\theta, w < 0)$$

positively weighted subset

negatively weighted subset

The mixture components are determined by the sign of the weights

# Breaking up the Learning Task

Decompose the quasi-pdfs into signed mixture models

$$q(x|\theta) = c(\theta)p(x|\theta, w \geq 0) - (c(\theta) - 1)p(x|\theta, w < 0)$$

The mixture coefficient is not a probability  $c(\theta) \geq 1$

Estimator  $\hat{c} = \frac{\sum_{i, w_i \geq 0} w_i}{\sum_i w_i}$

# Ratio of Signed Mixtures Model (RoSMM)

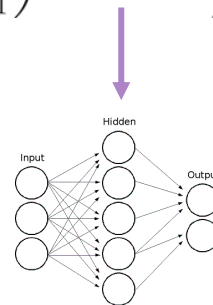
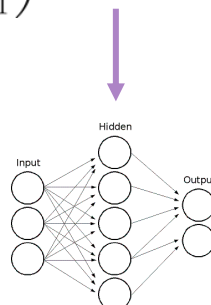
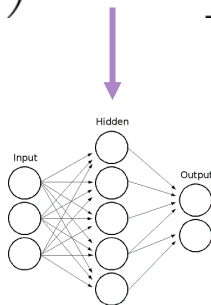
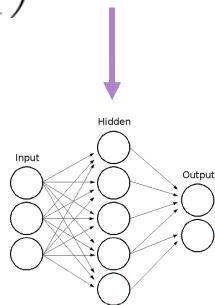
Decompose the quasi-pdfs into signed mixture models

$$q(x|\theta) = c(\theta)p(x|\theta, w \geq 0) - (c(\theta) - 1)p(x|\theta, w < 0)$$

The density ratio can be decomposed into four “sub-ratios”

$$\frac{q(x|\theta_1)}{q(x|\theta_0)} = \left[ \left( \frac{c_0}{c_1} \right) [r_{++}(x)]^{-1} + \left( \frac{1-c_0}{c_1} \right) [r_{-+}(x)]^{-1} \right]^{-1} + \left[ \left( \frac{c_0}{1-c_1} \right) [r_{+-}(x)]^{-1} + \left( \frac{1-c_0}{1-c_1} \right) [r_{--}(x)]^{-1} \right]^{-1}$$

Four  
classifiers

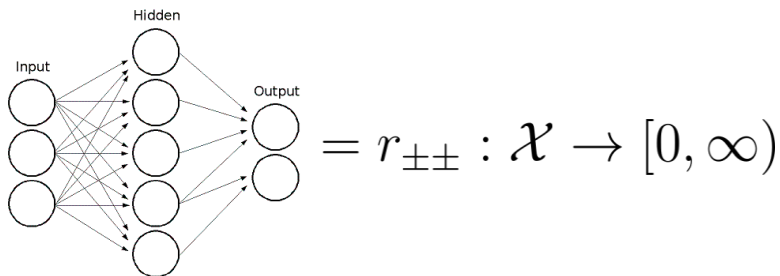


Combine the four “sub-ratios” to get a model for  $r(x|\theta)$

# RoSMM Models

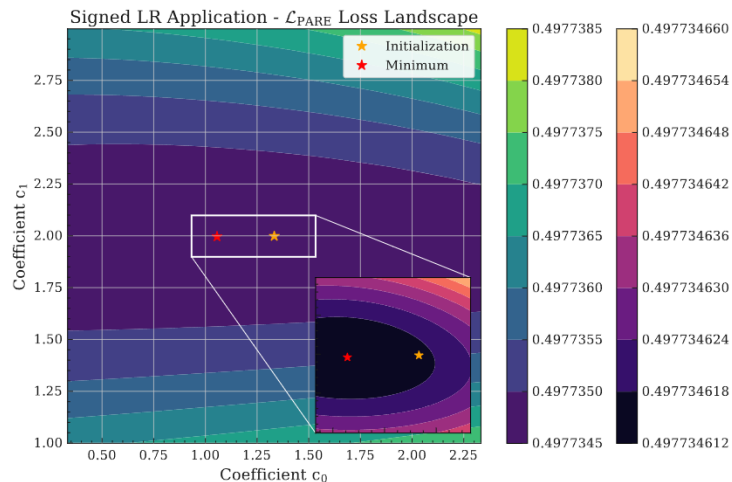
More negative weights is better!

1. We train four different carl NNs on the different disjoint nonnegative subsets of the data



2. Optimize the coefficients  $c_i \in \mathbb{R}$

*This can be done using the new loss function introduced earlier*



# Improved Performance vs Variance

Sample data from a toy probability distribution,  
but assign artificial weights with a controlled variance

$$W \sim 1 + \frac{\sigma_w}{\sqrt{\eta(1-\eta)}} (\eta - \text{Bern}(\eta))$$

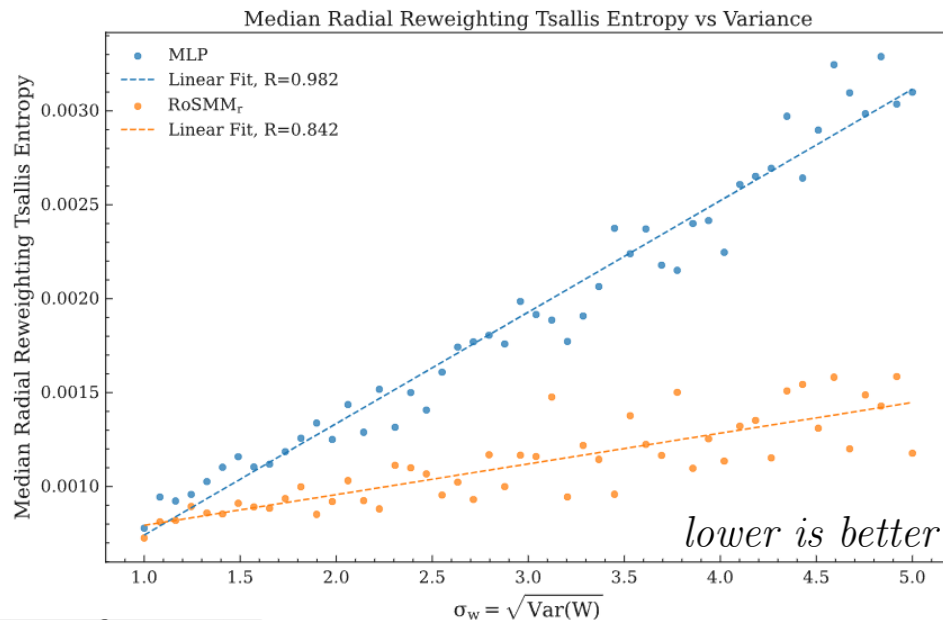
$$\mathbb{E}[W] = 1$$

$$\text{Var}(W) = \sigma_w^2$$

$$P(W < 0) = \eta$$

Compare the performance of each  
model as a function of the variance

Results were independent of the fraction of negative  
weights present (assuming sufficient sample sizes)



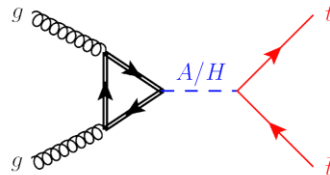
# Which Approach Should I Use?

	Solution 1 – New Loss/Classifier	Solution 2 – RoSMM
Can learn probabilistic density ratios?	Yes	Yes
Can learn quasiprobabilistic density ratios?	Yes	Yes
Model Complexity	One classifier	Two or four classifiers trained separately
Data requirements	None	Enough negatively weighted events to train a model
Variance	No change	Reduced

The best approach to use will likely depend on the number of negatively weighted events present and the variance of the weights

# Applications

# Search for



When interference is non-negligible, quasiprobabilities can arise

$$\mu S + \sqrt{\mu} I + B = (\mu - \sqrt{\mu}) \boxed{S} + \sqrt{\mu} \boxed{S + I} + \boxed{B}$$

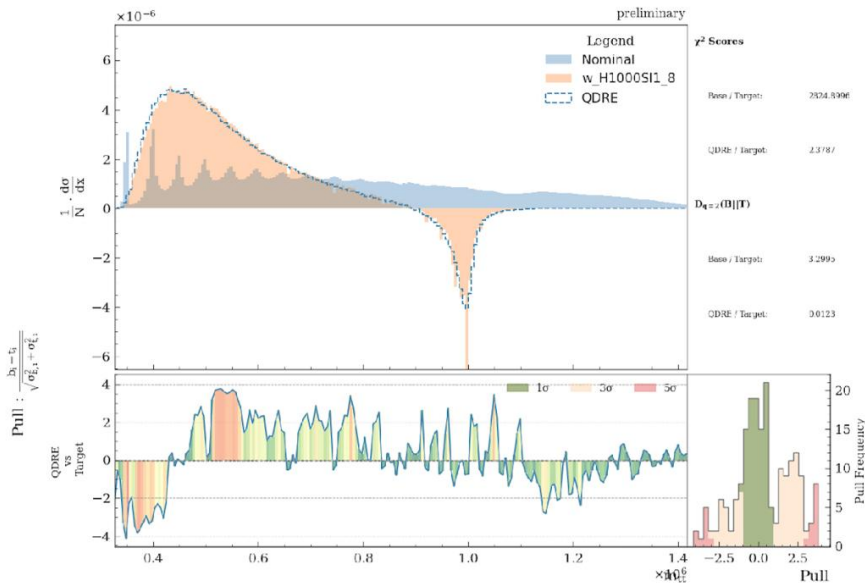
nonnegative pdf

quasi-pdf

Density ratios can be used as an importance weight to sample from quasi-pdfs

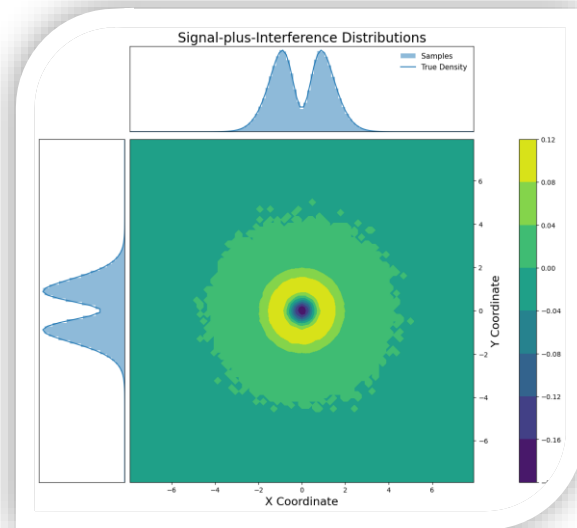
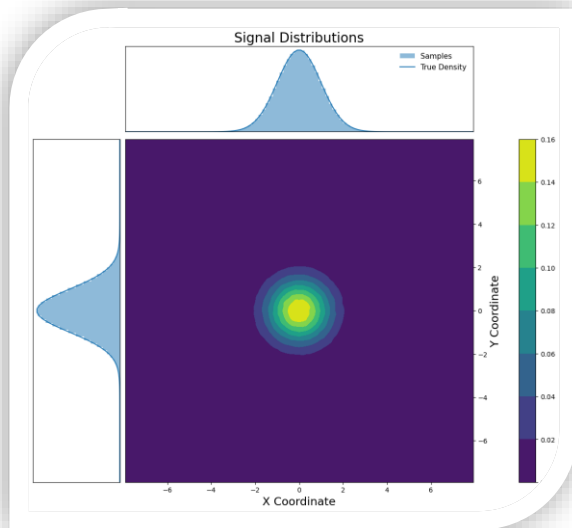
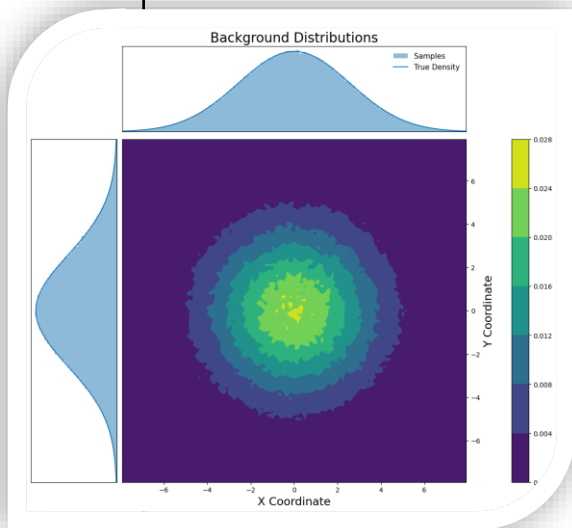
$$\hat{r}(x|\theta) \approx \frac{q_{S+I}(x|\theta)}{p_S(x|\theta)}$$

We can train a model to learn this parameterized reweighting



# SBI with Explicit Interference

Consider a toy analogue of the previous setup example, where background, signal, and signal+interference are each described separately



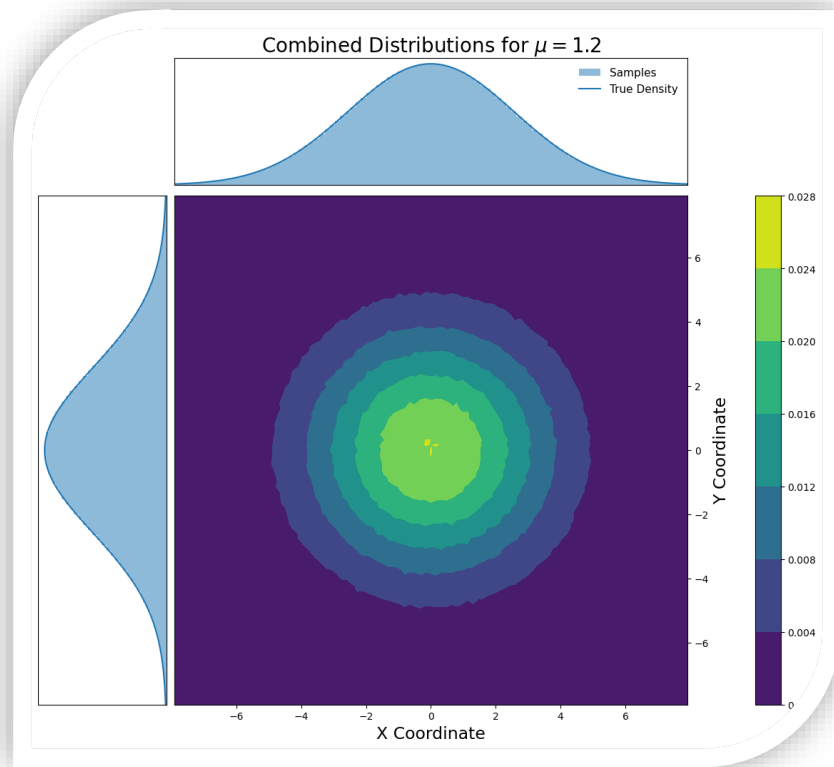
**this goes negative!**

# SBI with Explicit Interference

Choose a signal strength of  $\mu = 1.2$   
and  $\sigma_S/\sigma_B \approx 1/1600$

For inference, generate:

- 9995240 background events
- 653 signal events
- 4105 signal + interference events



# SBI with Explicit Interference

Parameterize the density in terms of background, signal, and interference

$$p(x|\mu) = \frac{\sigma_B p_B(x) + \sqrt{\mu}\sigma_{SI} q_{SI}(x) + (\mu - \sqrt{\mu})\sigma_S p_S(x)}{\sigma_B + \sqrt{\mu}\sigma_{SI} + (\mu - \sqrt{\mu})\sigma_S}$$

nonnegative pdf

quasi-pdf

Choose a reference distribution (e.g. background/SM)

$$\frac{p(x|\mu)}{p(x|\mu=0)} = (\sigma_B + \sqrt{\mu}\sigma_{SI} + (\mu - \sqrt{\mu})\sigma_S)^{-1} \left( \sigma_B + \sqrt{\mu}\sigma_{SI} \frac{q_{SI}(x)}{p_B(x)} + (\mu - \sqrt{\mu})\sigma_S \frac{p_S(x)}{p_B(x)} \right)$$

Now we need to estimate a ratio instead of a density

$$r(x|\mu) = (\sigma_B + \sqrt{\mu}\sigma_{SI} + (\mu - \sqrt{\mu})\sigma_S)^{-1} (\sigma_B + \sqrt{\mu}\sigma_{SI} r_{SI}(x) + (\mu - \sqrt{\mu})\sigma_S r_S(x))$$

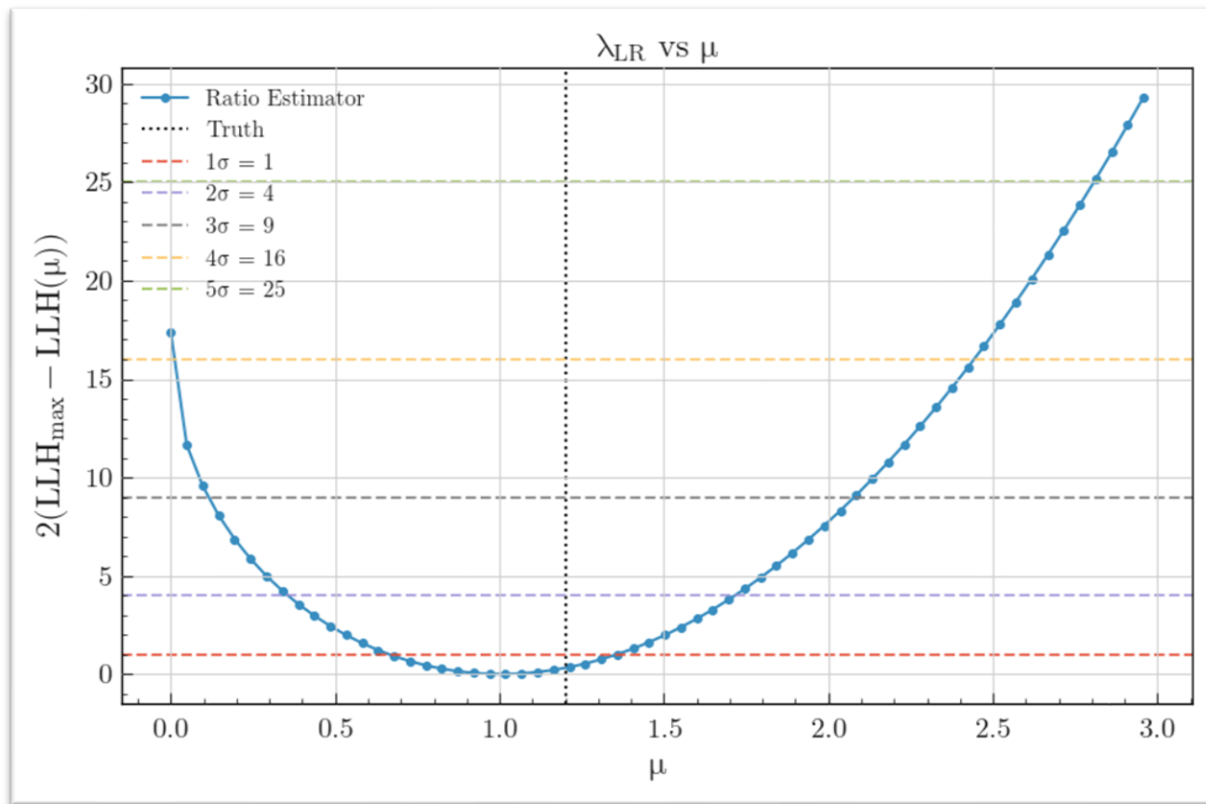
And we can get a likelihood ratio test statistic

$$\lambda_{LR} = -2 [\ln r(x|\mu) - \ln r(x|\hat{\mu})]$$

**Train two ratio estimators**

# SBI with Explicit Interference

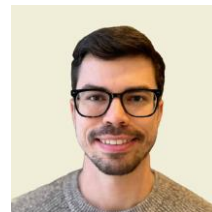
Look at the confidence intervals on  $\mu$



# SUMMARY

- **Introduced neural likelihood ratio estimation methods** and demonstrated their limitations
- **Extended neural classification-based density ratio estimation** to be compatible with quasiproability distributions [\[2512.19913\]](#)
- **Developed a new model architecture for density ratio estimation** that leverages negative weights to reduce the model training variance [\[arxiv:2410.10216\]](#)
- **Applied to modelling BSM distributions** with explicit interference (see my thesis)
- **Demonstrated SBI** with explicit interference modeling on a toy dataset

M. Drnevich



S. Jiggins



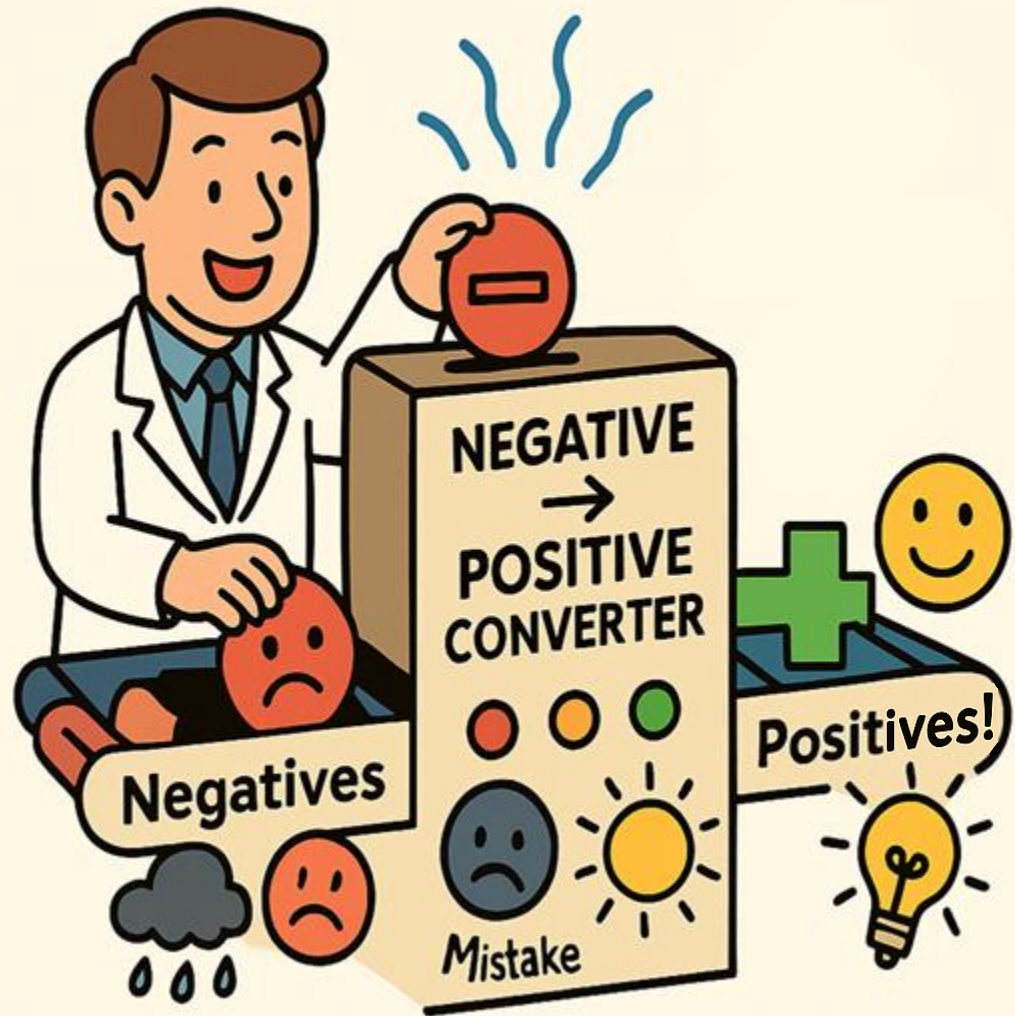
K. Cranmer



J. Katzy



THANK YOU



BACKUP

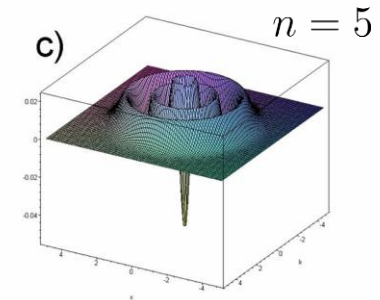
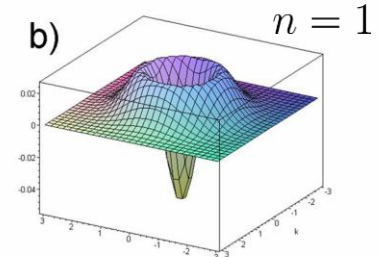
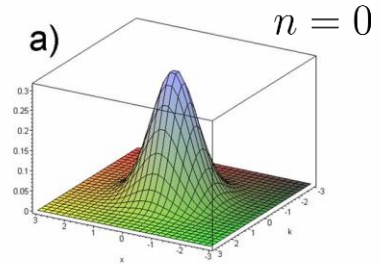
# Quasiprobabilities & Quantum Physics

Quasiprobability distributions allow for negative probabilities but maintain proper expectation values of observables

$$\langle \psi | \hat{G} | \psi \rangle = \text{Tr}(\hat{\rho} \hat{G}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \underbrace{W(x, p)}_{\text{Quasi-pdf}} g(x, p) dx dp$$

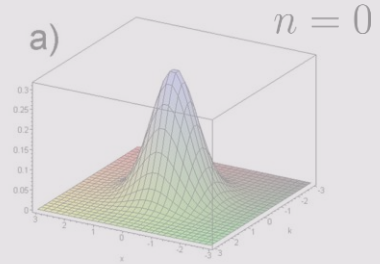
They can also be used to represent corrections to existing models

$$\int_{\Omega} |\mathcal{M}|^2 d\Phi(\mathbf{x}) = \int_{\Omega} \underbrace{(|\mathcal{M}_{SM}|^2)}_{\text{pdf}} + \underbrace{(|\mathcal{M}_{BSM}|^2 + 2\text{Re}(\mathcal{M}_{SM} \mathcal{M}_{BSM}^*))}_{\text{Quasi-pdf}} d\Phi(\mathbf{x})$$



# Quasiprobabilities & Quantum Physics

Quasiprobability distributions yield proper expectation values with respect to the weights of the distribution



$\langle \psi | \hat{A} \psi \rangle = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots$

*"It is usual to suppose that, since the probabilities of events must be positive, a theory which gives negative numbers for such quantities must be absurd ... By discussing a number of examples, I hope to show that they are entirely rational of course, and that their use simplifies calculation and thought in a number of applications."*

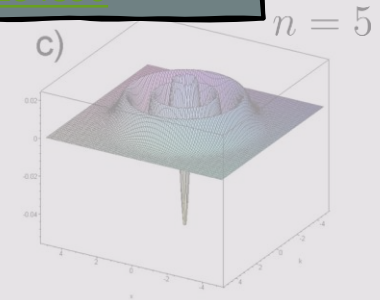
- Richard Feynman, Negative Probability <https://cds.cern.ch/record/154856>



Th  
to existing models

$$|\mathcal{M}|^2 = |\mathcal{M}_{SM}|^2 + |\mathcal{M}_{BSM}|^2 + \underbrace{\mathcal{M}_{SM}^* \mathcal{M}_{BSM} + \mathcal{M}_{SM} \mathcal{M}_{BSM}^*}_{\text{Quasi-pdf}}$$

Quasi-pdf



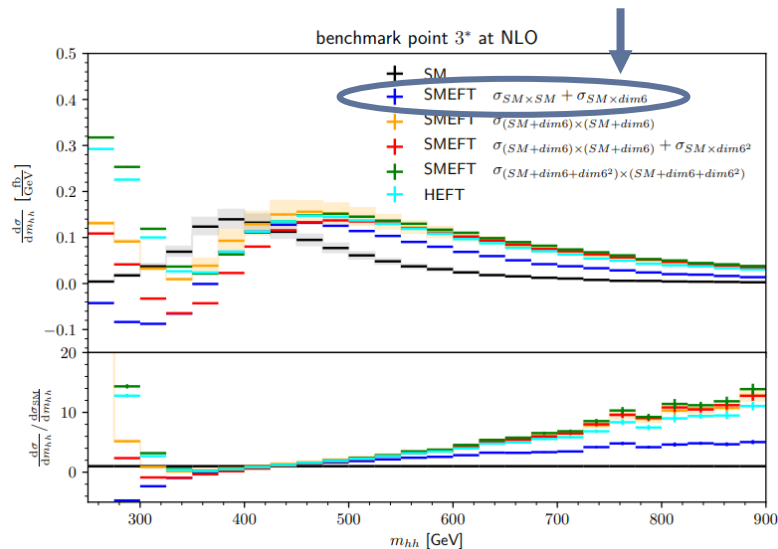
# SMEFT: $gg \rightarrow hh$

- Data is generated using Powheg + Pythia
- Classifying/reweighting between Standard Model and SM + EFT
- The truncation chosen has a quasiprobability distribution

$$\begin{aligned}
 \mathcal{M} = & \text{[Diagrams: SM and dim-6 operators]} \\
 & + \text{[Diagrams: dim-8 operators]} + \dots \\
 = & \mathcal{M}_{\text{SM}} + \mathcal{M}_{\text{dim6}} + \mathcal{M}_{\text{dim6}^2},
 \end{aligned}$$

[arXiv:2204.13045]

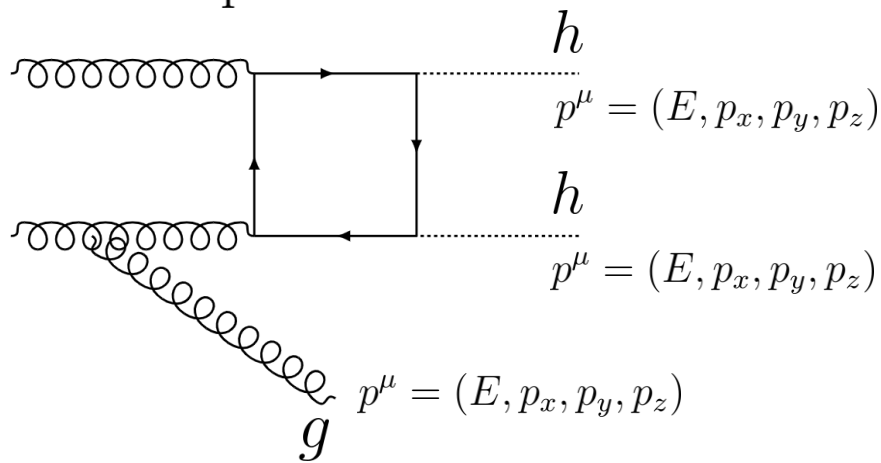
WE USE THIS TRUNCATION FOR THE ALTERNATIVE/TARGET



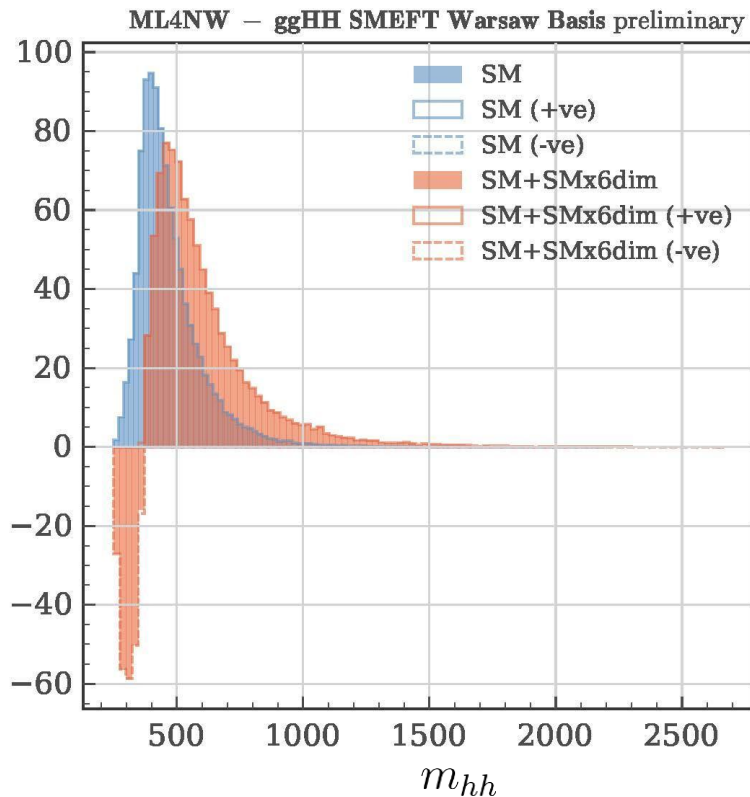
benchmark (* = modified)	$C_{H,\text{kin}}$	$C_H$	$C_{uH}$	$C_{HG}$	$\Lambda$
SM	0	0	0	0	1 TeV
1*	4.95	-6.81	3.28	0	1 TeV
3*	13.5	2.64	12.6	0.0387	1 TeV
6*	0.561	3.80	2.20	0.0387	1 TeV

# SMEFT: $gg \rightarrow hh$ at NLO

For example:



We consider the final state  
4-momentum distributions  $q(p^\mu | \text{Higgs})$



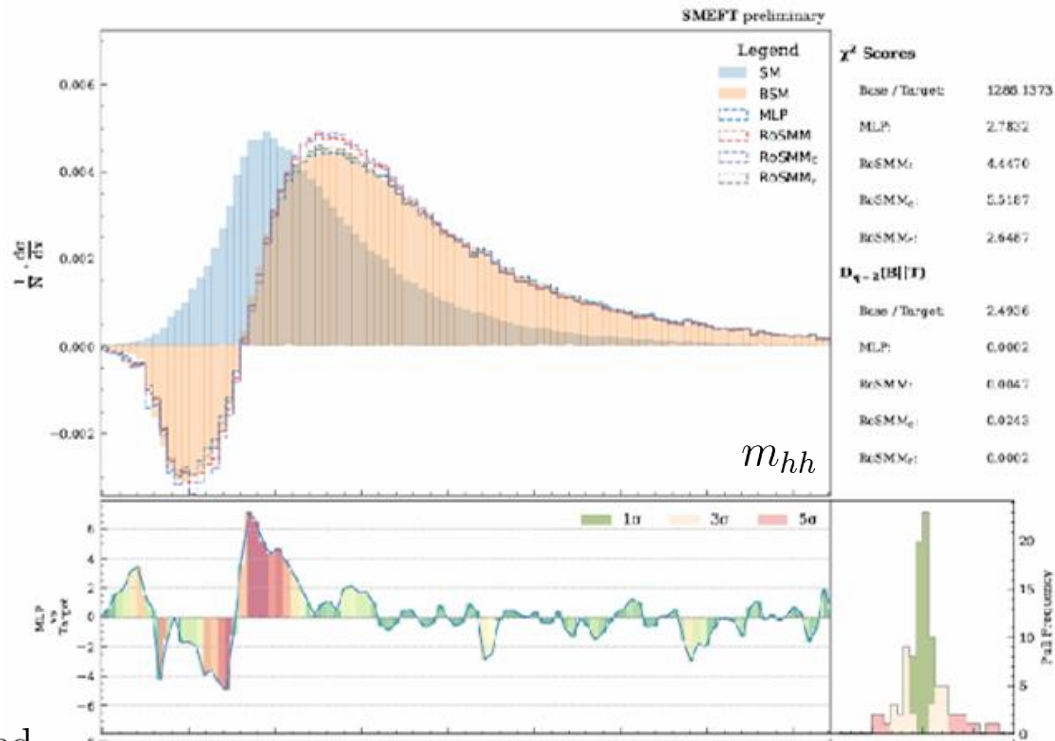
# SMEFT RESULTS

Training is performed on the final state 4-momentums

$$\left\{ p_T^{(i)}, \eta^{(i)}, \phi^{(i)}, m^{(i)} \right\}$$

Performance is evaluated on the reconstructed di-Higgs mass distribution

The density ratio estimate is used as an importance sampling weight mapping source ( $SM$ ) to target ( $SM + SM \times dim6$ )



$$\mathbb{E}_{x \sim q_{BSM}(x)} [f(x)] = \mathbb{E}_{x \sim p_{SM}(x)} [f(x) \hat{r}(x)]$$

# CAUTIONARY NOTE

Regions of phase space with negative density can be “hiding” in higher dimensions

These may not be visible in lower dimensional projections or observables

Some diagnostic tools are in development so stay tuned!

