

# HistFactory v2:

## Statistical modeling using Gaussian Process Regression

Simulation-Based Inference Blueprint Workshop

Kyle Cranmer, Matthew Feickert, Lukas Heinrich, Alexander Held, Jay Sandesara



**WISCONSIN**  
UNIVERSITY OF WISCONSIN-MADISON

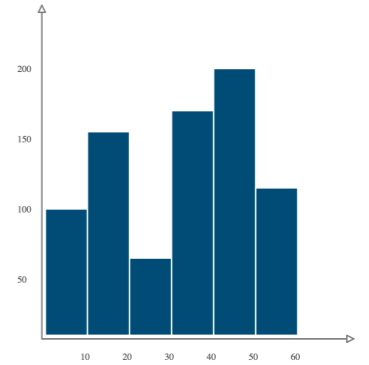


# Introduction

- The talk will cover an exciting new set of ideas that allow robust modeling in high-dimensional systematic uncertainty space with minimal assumptions and expressive models.
- The ideas will be presented in the context of traditional binned template analysis, but they apply also to Simulation-Based Inference - as described in the last slides.
- This is ongoing work with publication to be out soon, along with an implementation in ROOT, pyhf and also the IRIS-HEP SBI toolkit!
- The full, more technical talk was presented at the Statistical Ecosystem Blueprint workshop on Tuesday.

# Binned HistFactory Model

# The Binned Probability Model



The binned template model we are familiar with:

$$P(n, a|\eta, \chi) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb} | \nu_{cb}(\eta, \chi))}_{\text{bin-by-bin Poisson PDFs}} \cdot \underbrace{\prod_{\chi_p \in \chi} f_{\chi_p}(a_{\chi_p} | \chi_p)}_{\text{Constrain terms}}$$

$n \rightarrow$  Observed events

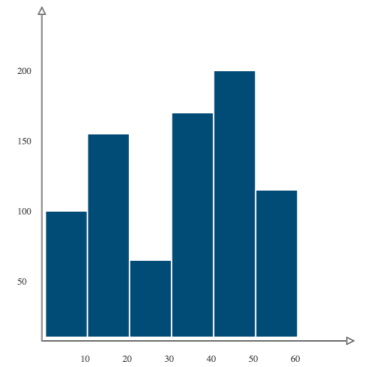
$f_{\chi_p} \rightarrow$  Constrain on each parameter  $\chi_p$

$\nu \rightarrow$  Expected events

$a_{\chi_p} \rightarrow$  associated auxiliary data

	Parameter	HistFactory modifiers
Unconstrained params	$\eta$	NormFactor ShapeFactor
Constrained params $\chi$	$\xi$	StatError ShapeSys
	$\alpha$	OverallSys HistoSys

# The Binned Probability Model



The binned template model we are familiar with:

$$P(n, a|\eta, \chi) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb} | \nu_{cb}(\eta, \chi)) \cdot \prod_{\chi_p \in \chi} f_{\chi_p}(a_{\chi_p} | \chi_p)$$

Key piece to estimate

$$\begin{aligned} \nu_{cb}(\eta, \chi) &= \sum_{s \in \text{samples}} \nu_{scb}(\eta, \chi = (\xi, \alpha)) \\ &= \sum_{s \in \text{samples}} \left[ \prod_{\eta, \xi} \kappa_{scb}(\eta, \xi) \right] \cdot \Delta_{scb}(\alpha) \cdot \nu_{scb}^0 \end{aligned}$$

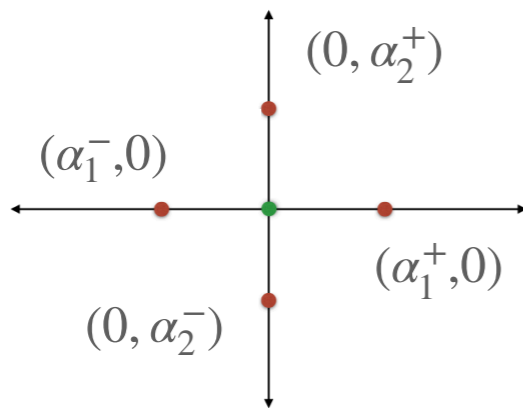
Norm factors  
e.g. POIs, lumi, MC stat  
(known a-priori parametric dependence)

Need to estimate  
(no known a-priori dependence)

# Traditional HistFactory Model

$$\nu_{cb}(\eta, \chi) = \sum_{s \in \text{samples}} \left[ \prod_{\eta, \xi} \kappa_{scb}(\eta, \xi) \right] \cdot \underbrace{\Delta_{scb}(\alpha)}_{\text{blue bracket}} \cdot \nu_{scb}^0$$

$$\Delta_{scb}(\alpha) = \frac{\nu_{scb}(\alpha)}{\nu_{scb}^0}$$



Example

anchor points for  $\alpha \in \mathbb{R}^2$

Known at specific "anchor points"  $\alpha^{(i)}$   
where simulations are available.

Need model for continuous dependence  
across  $\alpha$ -space.

**Challenges:**

**High-dim space**  $\alpha \in \mathbb{R}^{100-1000}$

**No known parametric dependence**

# Traditional HistFactory Model

$$\nu_{cb}(\eta, \chi) = \sum_{s \in \text{samples}} \left[ \prod_{\eta, \xi} \kappa_{scb}(\eta, \xi) \right] \cdot \Delta_{scb}(\alpha) \cdot \nu_{scb}^0$$

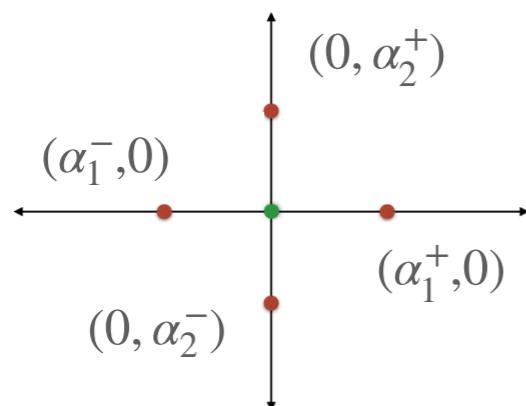
Traditional HistFactory model relies on two key assumptions:

**Assumption 1: Factorization**

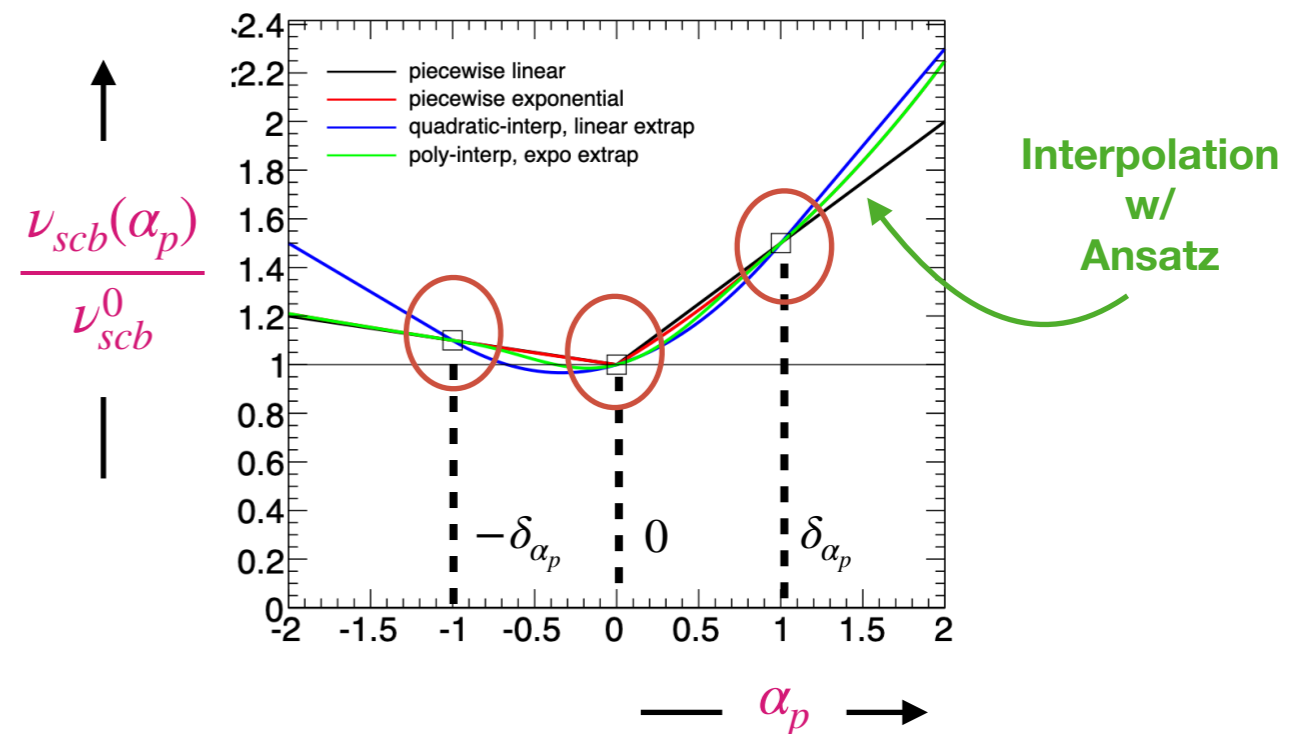
**Assumption 2: Parametric Ansatz**

The impact factorize:

$$\Delta_{scb}(\alpha) = \frac{\nu_{scb}(\alpha)}{\nu_{scb}^0} = \prod_p \frac{\nu_{scb}(\alpha_p)}{\nu_{scb}^0}$$



Available simulations at  $\alpha_p = \{0, \pm \delta_{\alpha_p}\}$

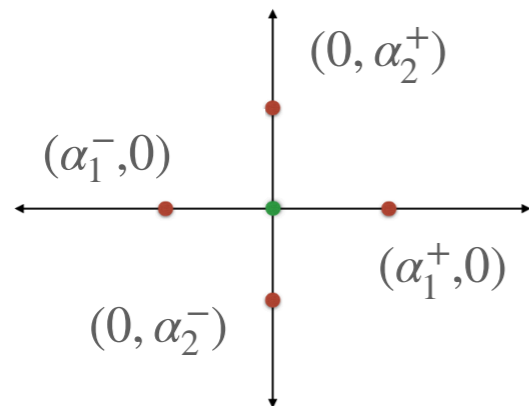


# Traditional HistFactory Model

## Assumption 1: Factorization

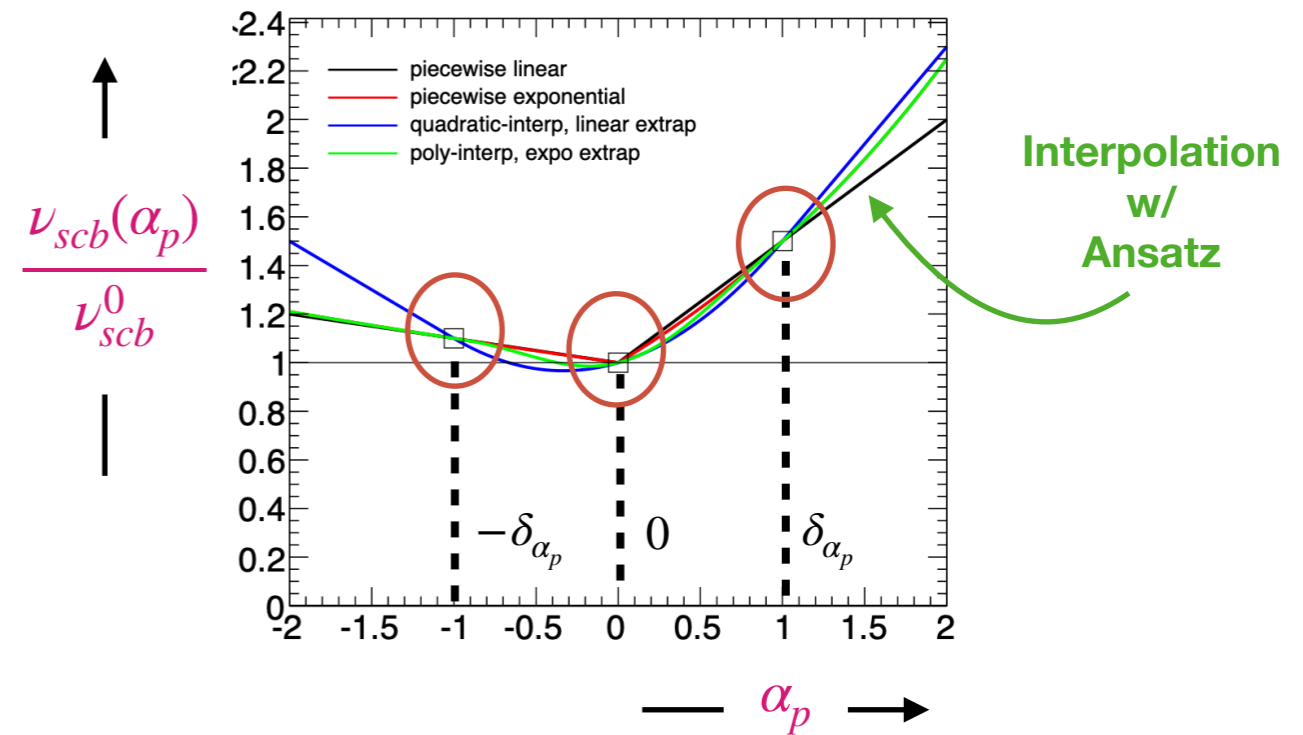
The impact factorize:

$$\Delta_{scb}(\alpha) = \frac{\nu_{scb}(\alpha)}{\nu_{scb}^0} = \prod_p^{N_{syst}} \frac{\nu_{scb}(\alpha_p)}{\nu_{scb}^0}$$



Available simulations at  $\alpha_p = \{0, \pm \delta_{\alpha_p}\}$

## Assumption 2: Parametric Ansatz



Example HistFactory model:

$$\Delta_{scb}(\alpha) = \prod_p I_{poly|exp}(\alpha; \mathbf{1}, \Delta_{scb}(\alpha_p^+), \Delta_{scb}(\alpha_p^-))$$

Factorization
Analytic ansatz

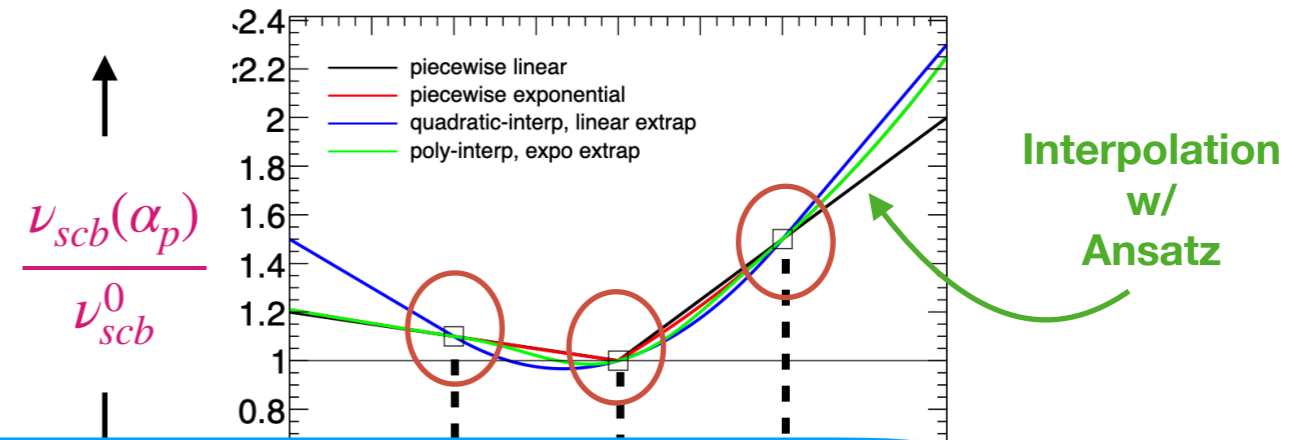
# Traditional HistFactory Model

## Assumption 1: Factorization

The impact factorize:

$$\Delta_{scb}(\alpha) = \frac{\nu_{scb}(\alpha)}{\nu_{scb}^0} = \prod_p^{N_{syst}} \frac{\nu_{scb}(\alpha_p)}{\nu_{scb}^0}$$

## Assumption 2: Parametric Ansatz



## Limitation:

The impact on final likelihood may not factorize

$$\Delta_{scb}(\alpha) = \prod_p^{N_{syst}} \frac{\nu_{scb}(\alpha_p)}{\nu_{scb}^0} \times \text{corrections?}$$

Model may not be expressive to cover more intricate setups

**HistFactory v2:**

**Modeling using Gaussian Process Regression**

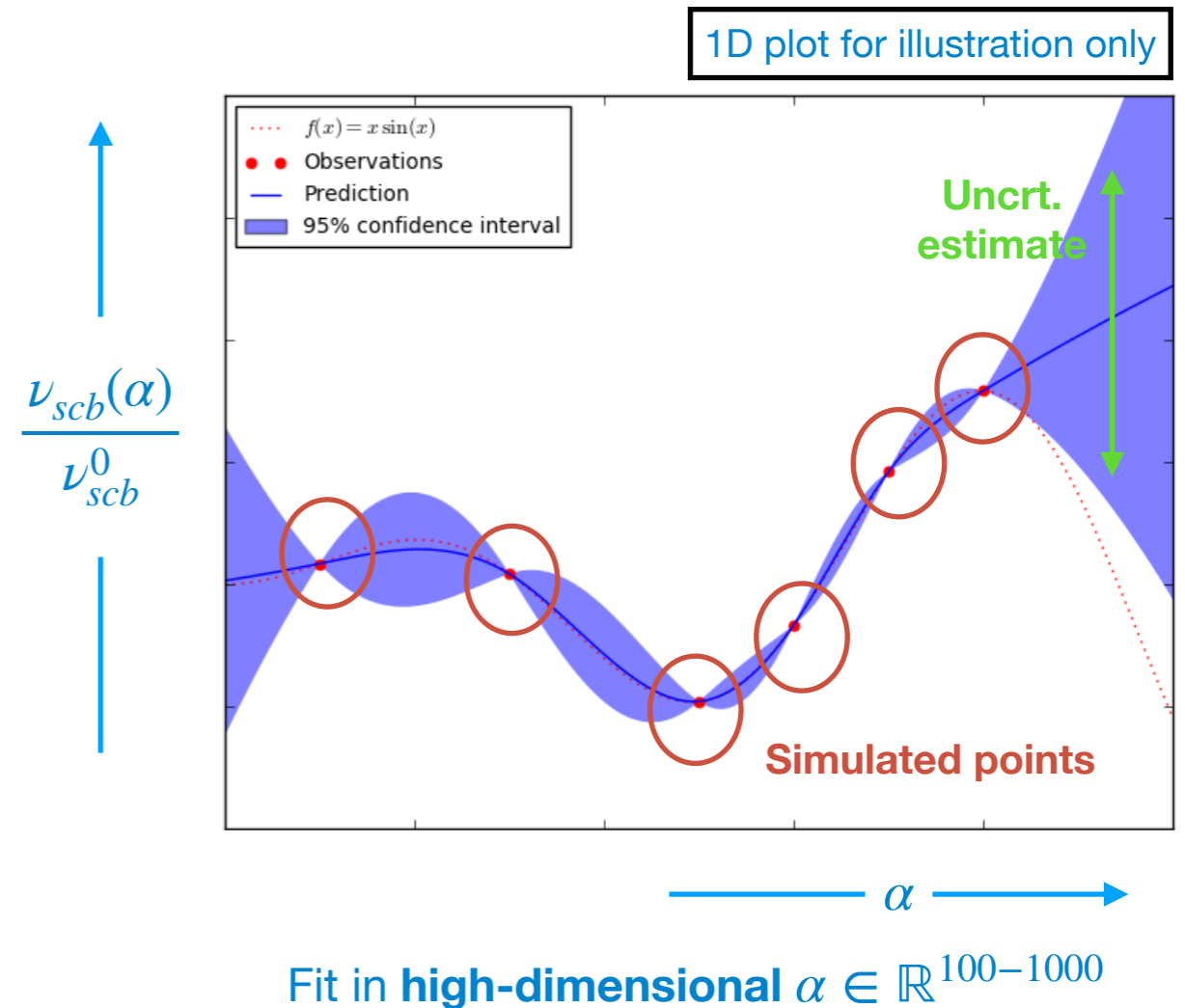
# Gaussian Process Regression

## Gaussian Processes:

Generalization of Gaussian *distributions* to an infinite-dimensional space of functions.

Completely specified by a mean function  $m(x)$ , and covariance function  $k(x, x')$

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')).$$



## Proposal:

Use GP regression to model  $\Delta_{scb}(\alpha)$  term in the likelihood model?

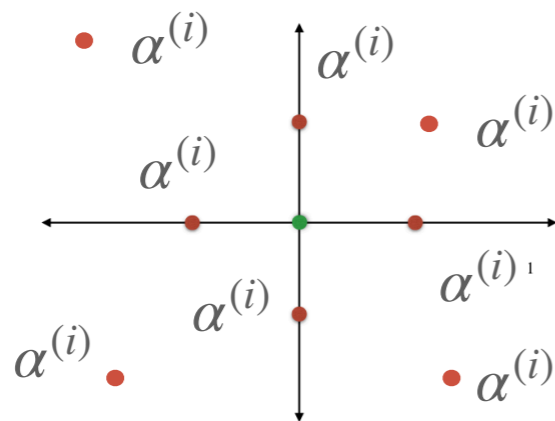
$$\nu_{cb}(\eta, \chi) = \sum_{s \in \text{samples}} \left[ \prod_{\eta, \xi} \kappa_{scb}(\eta, \xi) \right] \Delta_{scb}(\alpha) \nu_{scb}^0$$

# Inputs to Gaussian Processes

High-dimensional parameter points at which simulations are available

$$A_s^{sim} = \{\alpha^{(i)}\}$$

Not restricted by orthogonality



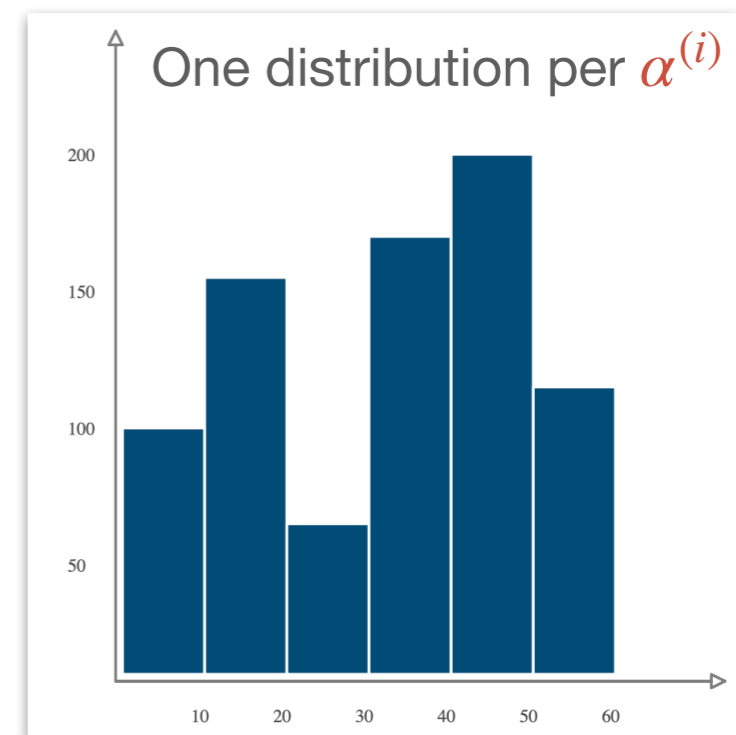
Example parameter space  $\{\alpha^{(i)}\}$  used in simulations

$$\alpha^{(i)} = \left( \alpha_1^{(i)}, \alpha_2^{(i)}, \dots, \alpha_{N_{syst}}^{(i)} \right) \in \mathbb{R}^{N_{syst}}$$

Predictions at simulation points

$$\Delta_{scb}(A_s^{sim}) = \left[ \frac{\nu_{scb}(\alpha^{(1)})}{\nu_{scb}^0}, \dots, \frac{\nu_{scb}(\alpha^{(N_{sim})})}{\nu_{scb}^0} \right]^T$$

$N_{sim}$  simulations or "anchor points"  $\alpha^{(i)}$



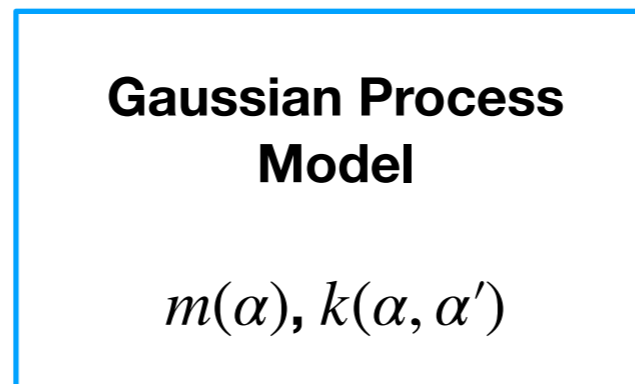
# Inputs to Gaussian Processes

High-dimensional parameter points  
at which simulations are available

$$A_s^{sim} = \{\alpha^{(i)}\}$$

Predictions at simulation points

$$\Delta_{scb}(A_s^{sim}) = \left[ \frac{\nu_{scb}(\alpha^{(1)})}{\nu_{scb}^0}, \dots, \frac{\nu_{scb}(\alpha^{(N_{sim})})}{\nu_{scb}^0} \right]^T$$



Parameter point for inference  
 $\alpha$

The prior choice  $m(\alpha), k(\alpha, \alpha')$   
characterizes the form of interpolation -  
think injection of "inductive bias"

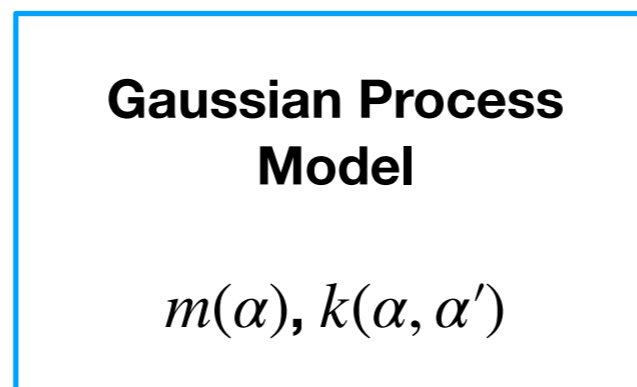
# Gaussian Process Regression

High-dimensional parameter points  
at which simulations are available

$$A_s^{sim} = \{\alpha^{(i)}\}$$

Predictions at simulation points

$$\Delta_{scb}(A_s^{sim}) = \left[ \frac{\nu_{scb}(\alpha^{(1)})}{\nu_{scb}^0}, \dots, \frac{\nu_{scb}(\alpha^{(N_{sim})})}{\nu_{scb}^0} \right]^T$$



Parameter point for inference  
 $\alpha$

$$p(\Delta'_{scb} \mid \alpha, A_s^{sim}, \Delta_{scb}(A_s^{sim})) = \mathcal{N}(\bar{\Delta}'_{scb}, \text{COV}(\Delta'_{scb}))$$



The final **posterior** prediction is a  
probability distribution of values

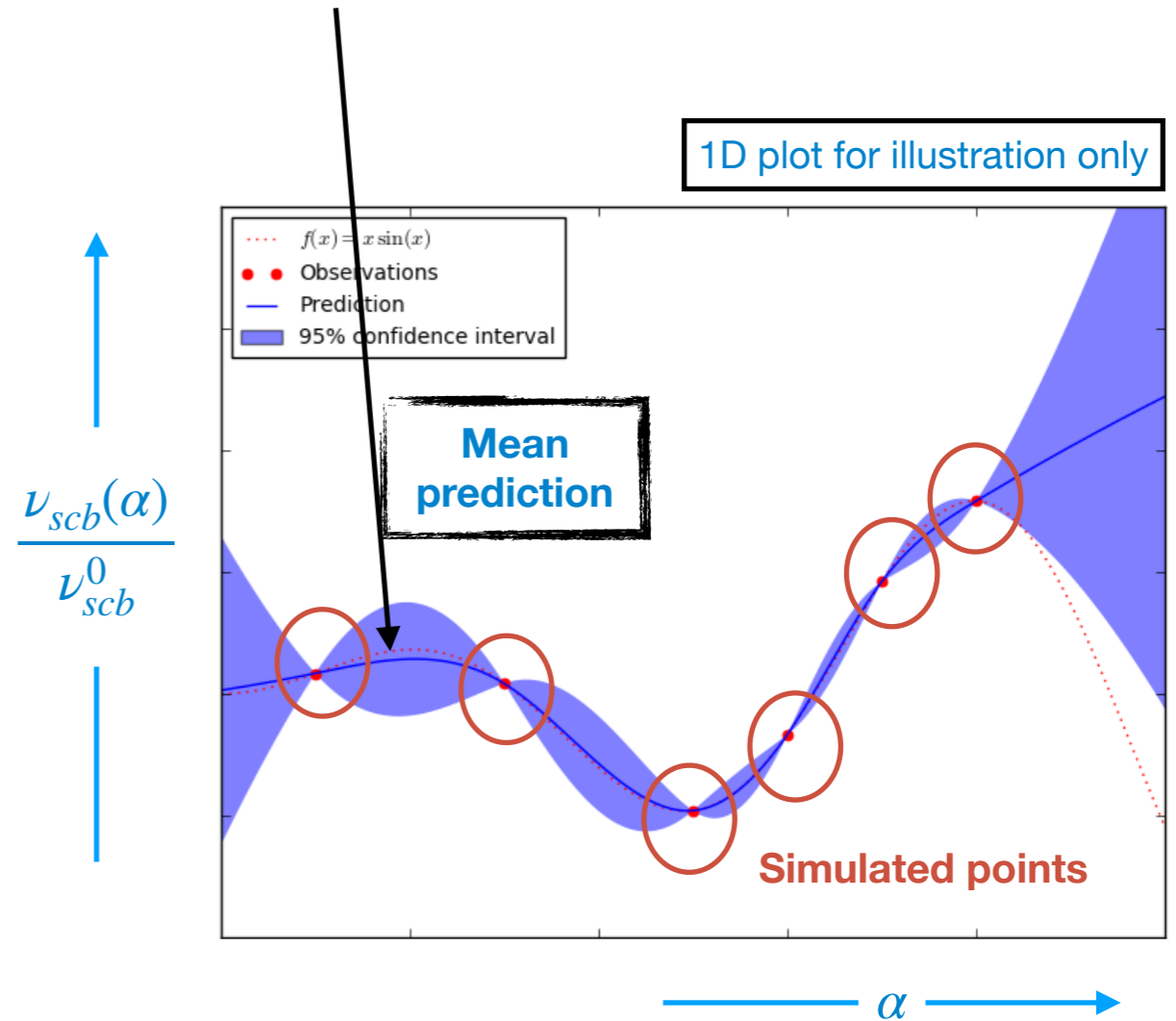
normal distribution with GP predicted  
mean and covariance kernel

# Gaussian Process Regression

$$p(\Delta'_{scb} \mid \alpha, \mathbf{A}_s^{\text{sim}}, \Delta_{scb}(\mathbf{A}_s^{\text{sim}})) = \mathcal{N}(\bar{\Delta}'_{scb}, \text{COV}(\Delta'_{scb}))$$

The final prediction is the posterior mean

$$\bar{\Delta}' = \Delta_{scb}(\alpha)$$



Fit in **high-dimensional**  $\alpha \in \mathbb{R}^{100-1000}$

# Gaussian Process Regression

$$p(\Delta'_{scb} \mid \alpha, \mathbf{A}_s^{\text{sim}}, \Delta_{scb}(\mathbf{A}_s^{\text{sim}})) = \mathcal{N}(\bar{\Delta}'_{scb}, \text{COV}(\Delta'_{scb}))$$

The final prediction is the posterior mean

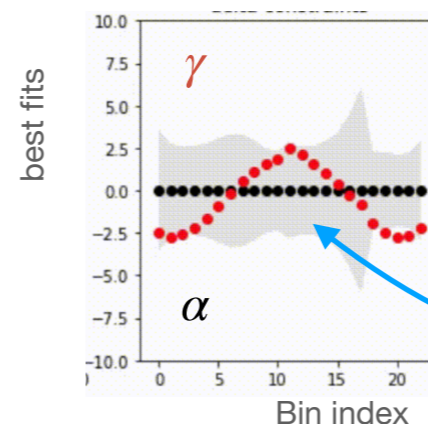
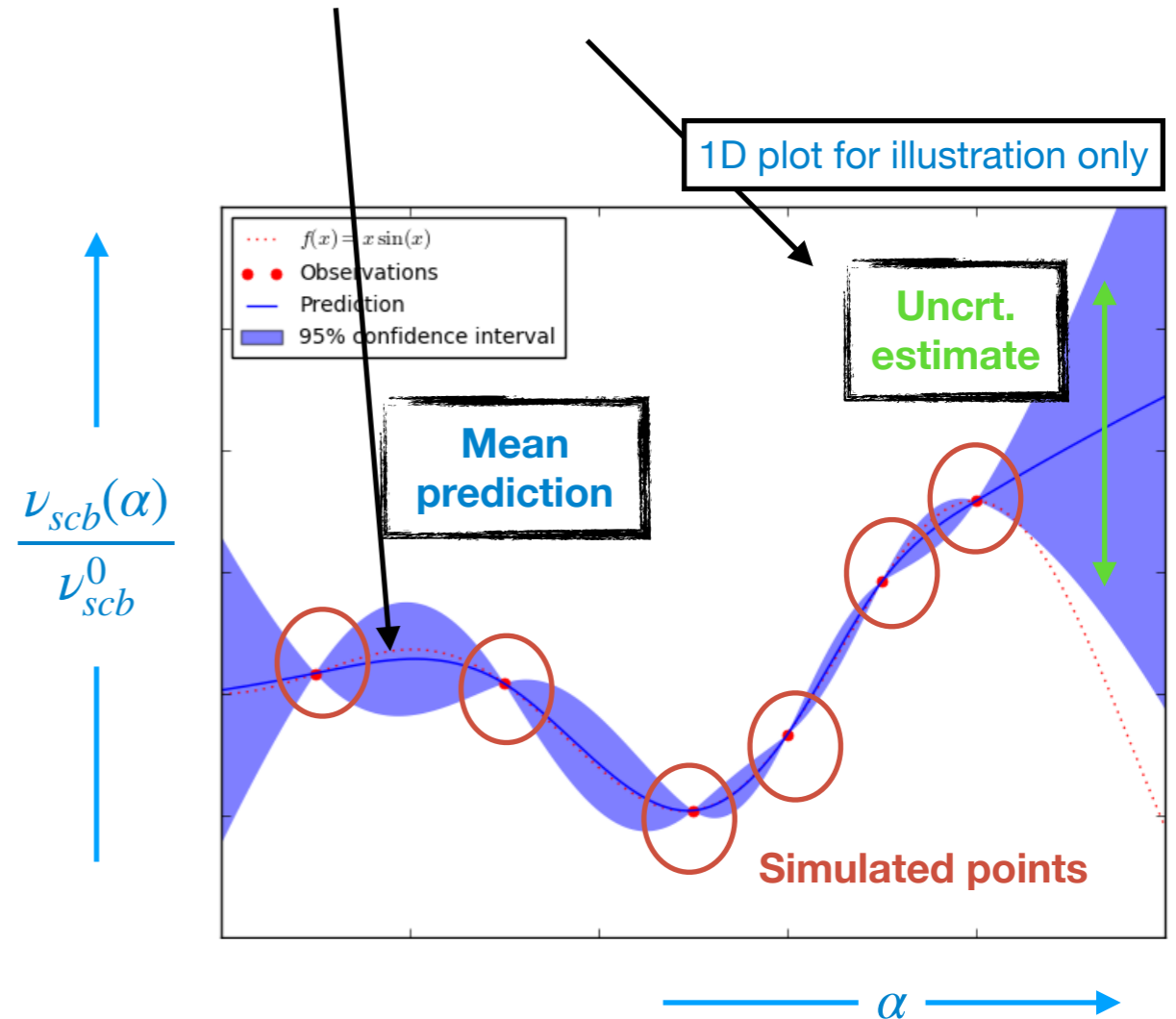
$$\bar{\Delta}' = \Delta_{scb}(\alpha)$$

Optional: Introduce additional bin parameter to account for uncertainties in morphing

$$\Delta_{scb}(\alpha) \cdot \kappa_{scb}(\xi = \gamma_b)$$

constrained by the covariance matrix predicted by GP regression

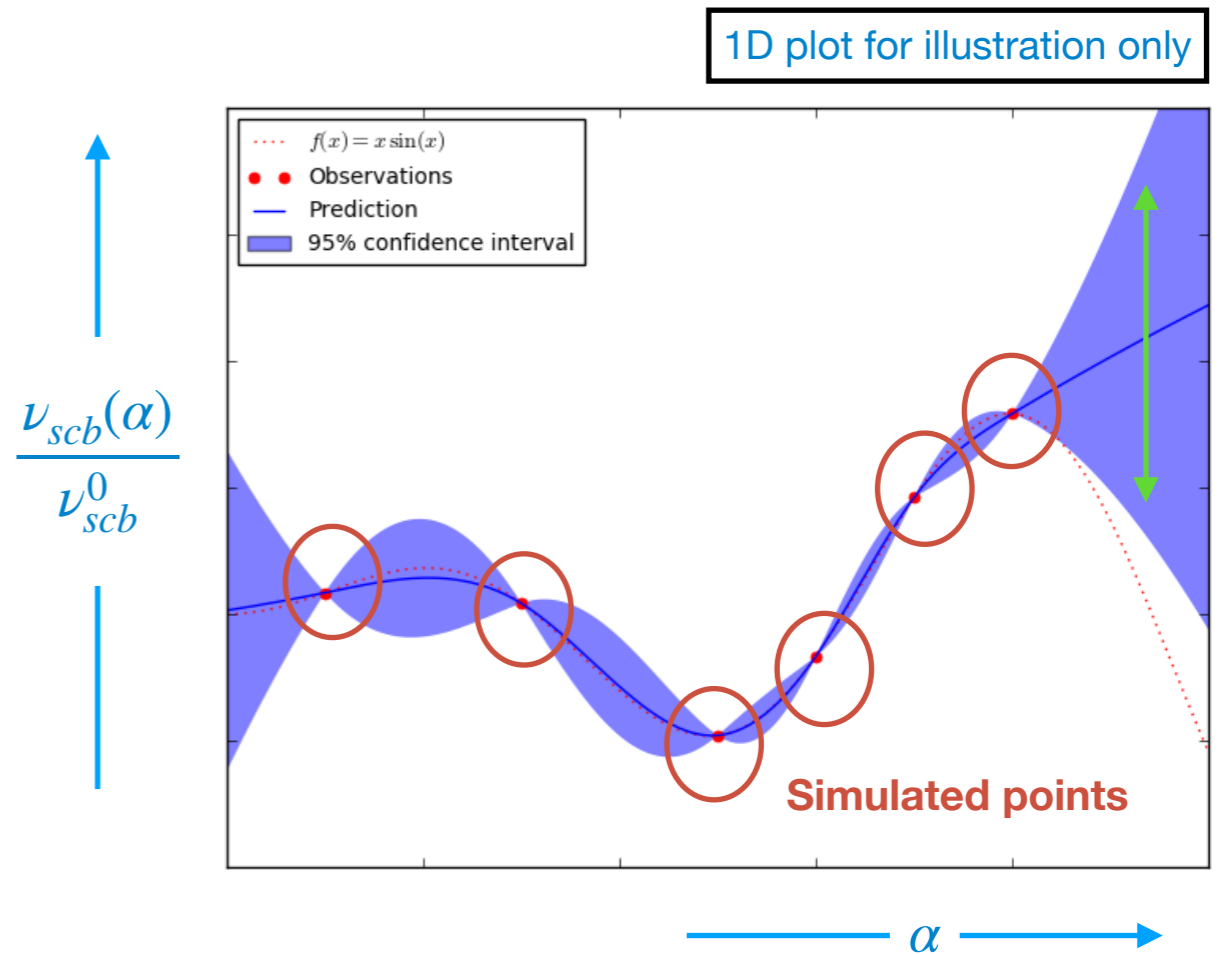
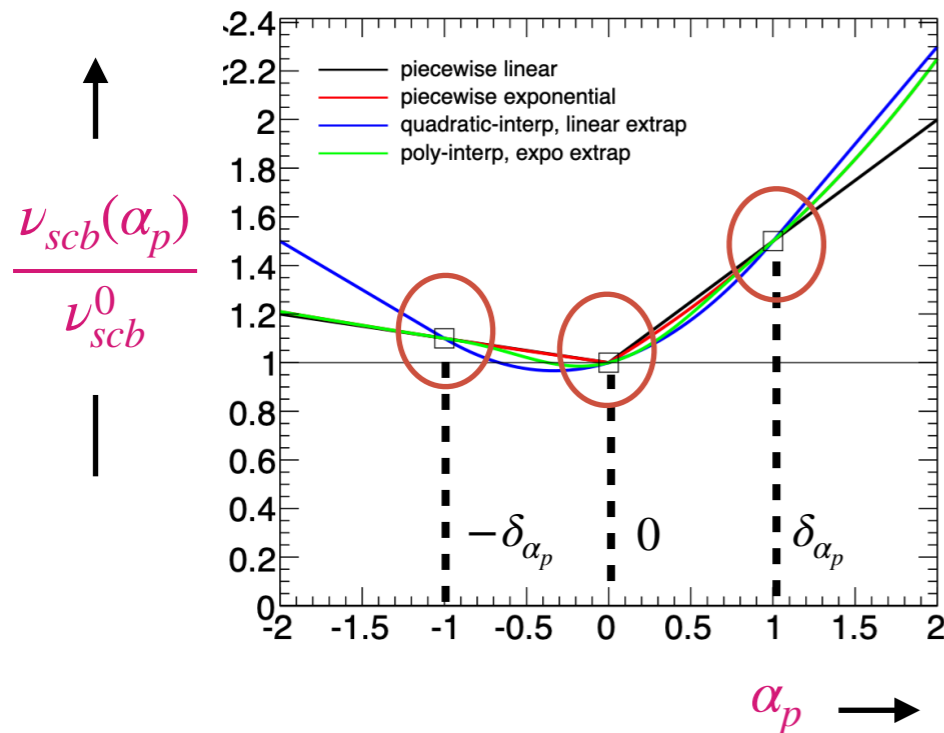
$$f_{\chi=\gamma}(\gamma_b, \alpha) = \text{Gaus}\left(1.0 \mid \gamma_b, \sqrt{\text{COV}(\Delta'_{scb})}\right)$$



We can now choose to propagate uncertainty on interpolation!

# Gaussian Process Regression

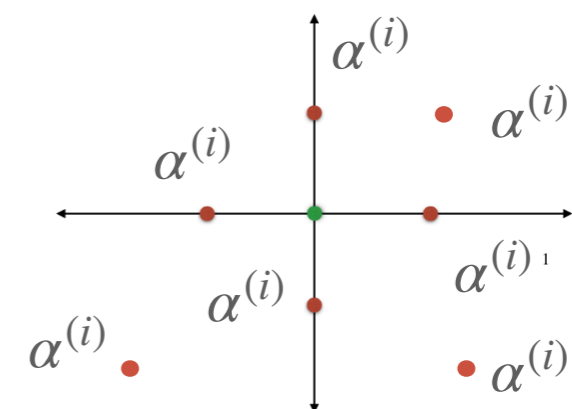
Think HistFactory-style interpolation:



Fit in high-dimensional  $\alpha \in \mathbb{R}^{100-1000}$

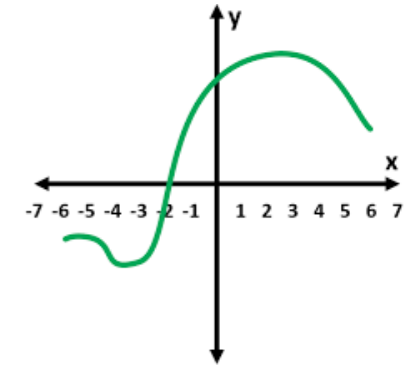
But with

**non-parametric model (more flexibility),**  
**high-dimensional fit (no factorization assumption),**  
**& with an uncertainty estimate (more robust)**



# The connection to Simulation-Based Inference

# The SBI Probability Model



The binned template model we are familiar with:

$$P(n, a|\eta, \chi) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb} | \nu_{cb}(\eta, \chi)) \cdot \prod_{\chi_p \in \chi} f_{\chi_p}(a_{\chi_p} | \chi_p)$$

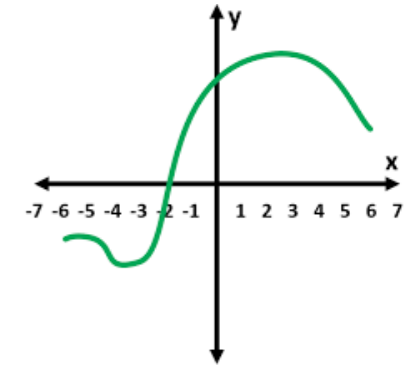
Generalization to unbinned probability density model

The unbinned SBI template model:

$$P(x|\eta, \chi) = \underbrace{\prod_{c \in \text{channels}} \prod_{e \in \text{events}_c} P(x_{ce} | \eta, \chi)}_{\text{event-by-event PDFs}} \cdot \underbrace{\prod_{\chi_p \in \chi} f_{\chi_p}(a_{\chi_p} | \chi_p)}_{\text{Constrain terms}}$$

(ignoring the extended term for simplicity)

# The SBI Probability Model



The unbinned SBI template model:

$$P(x|\eta, \chi) = \prod_{c \in \text{channels}} \prod_{e \in \text{events}_c} \boxed{P(x_{ce}|\eta, \chi)} \cdot \prod_{\chi_p \in \chi} f_{\chi_p}(a_{\chi_p}|\chi_p)$$

**HistFactory-style decomposition**

$$P(x_{ce}|\eta, \chi) = \frac{1}{\sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi)} \sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi) \cdot P_s(x_{ce} | \eta, \chi = (\xi, \alpha))$$

$$= \frac{1}{\sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi)} \sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi) \cdot \left[ \prod_{\eta, \xi} \kappa_{sc}(\eta, \xi) \right] \cdot \Delta_{sc}(x_{ce} | \alpha) \cdot P_s(x_{ce})$$

**Norm factors**

e.g. POIs, lumi, MC stat

(known a-priori parametric dependence)

**Need to estimate**

(no known a-priori dependence)

# The SBI Probability Model

Estimated using Neural Networks  
more in talk tomorrow

$$P(x_{ce}|\eta, \chi) = \frac{1}{\sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi)} \sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi) \cdot \left[ \prod_{\eta, \xi} \kappa_{sc}(\eta, \xi) \right] \cdot \Delta_{sc}(x_{ce} | \alpha) \cdot P_s(x_{ce})$$

In ATLAS, we used a HistFactory-style model

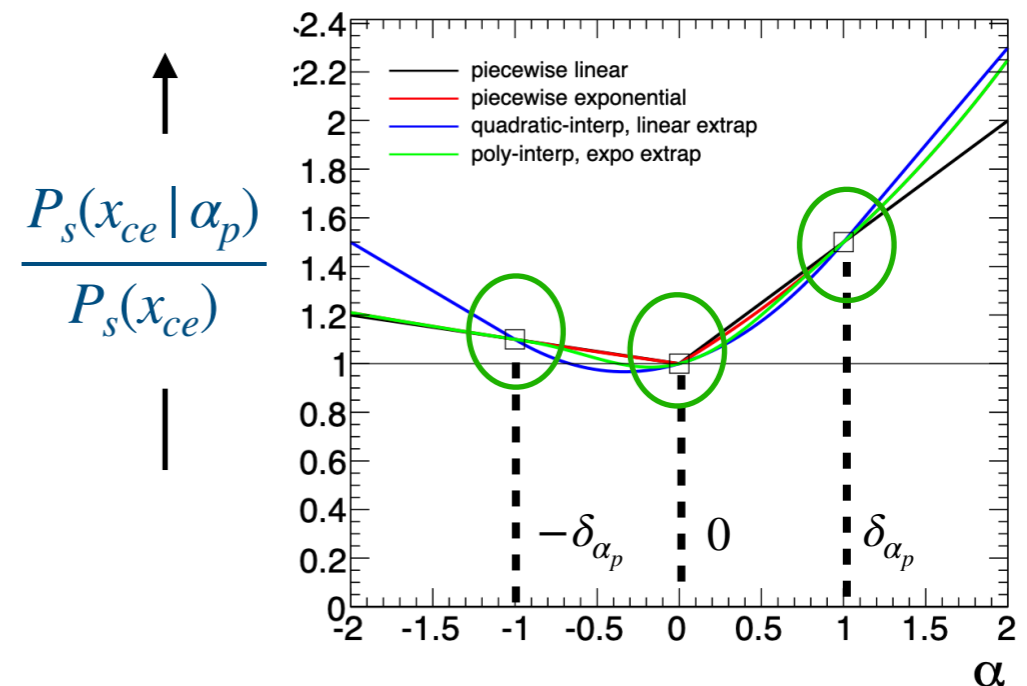
## Factorization assumption

$$\frac{P_s(x_{ce} | \alpha)}{P_s(x_{ce})} \approx \prod_p^{N_{\text{syst}}} \frac{P_s(x_{ce} | \alpha_p)}{P_s(x_{ce})}$$

$$\frac{P_s(x_{ce} | \alpha_p)}{P_s(x_{ce})} = \text{Interp} \left[ \underbrace{\frac{P_s(x_{ce} | + \delta_{\alpha_p})}{P_s(x_{ce})}}_{\text{Estimated using Neural Networks}}, \mathbf{1}, \underbrace{\frac{P_s(x_{ce} | - \delta_{\alpha_p})}{P_s(x_{ce})}}_{\text{Estimated using Neural Networks}} \right]$$

Estimated using Neural Networks

Interpolate/extrapolate from 3  
basis points -  
use ad-hoc parametric ansatz  
for continuous interpolation




# The SBI Probability Model

$$P(x_{ce}|\eta, \chi) = \frac{1}{\sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi)} \sum_{s \in \text{samples}} \nu_{sc}(\eta, \xi) \cdot \left[ \prod_{\eta, \xi} \kappa_{sc}(\eta, \xi) \right] \cdot \Delta_{sc}(x_{ce} | \alpha) \cdot P_s(x_{ce})$$

In ATLAS, we used a HistFactory-style model

[Rep. Prog. Phys. 88 067801](#)

EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)



Rep. Prog. Phys. 88 (2025) 067801  
DOI: 10.1088/1361-6633/add370


CERN-EP-2024-305  
June 16, 2025

**An implementation of neural simulation-based inference for parameter estimation in ATLAS**

The ATLAS Collaboration

[Rep. Prog. Phys. 88 057803](#)

EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

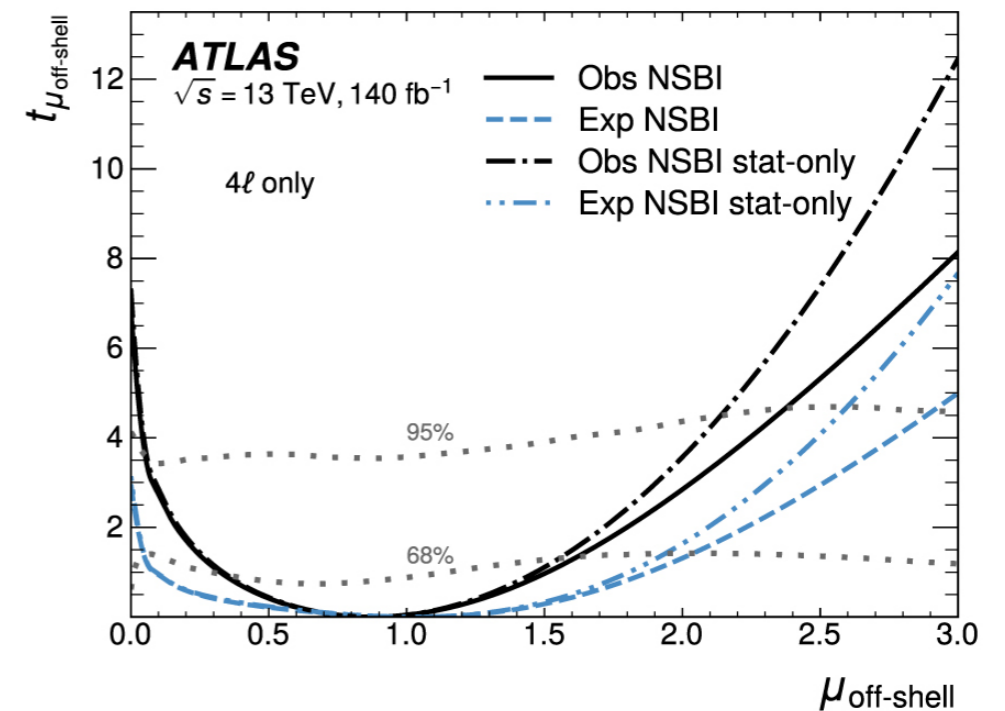


Rep. Prog. Phys. 88 (2025) 057803  
DOI: 10.1088/1361-6633/adcd9a

CERN-EP-2024-298  
May 23, 2025

**Measurement of off-shell Higgs boson production in the  $H^* \rightarrow ZZ \rightarrow 4\ell$  decay channel using a neural simulation-based inference technique in 13 TeV  $pp$  collisions with the ATLAS detector**

The ATLAS Collaboration



First ever measurement with real data using fully-differential multi-variate information

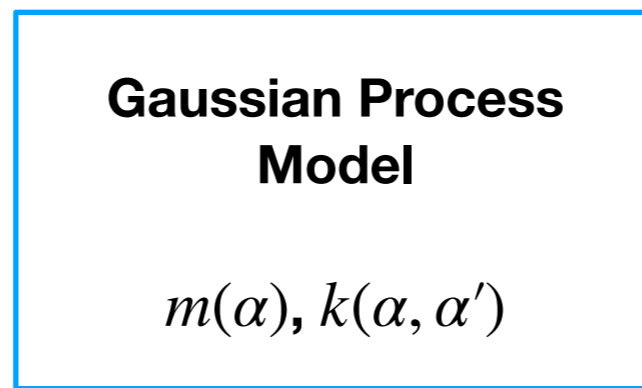
# Gaussian Process Regression for SBI?

High-dimensional parameter points at which simulations are available

$$A_s^{sim} = \{\alpha^{(i)}\}$$

Predictions at simulation points

$$\left[ \frac{P_s(x_{ce} | \alpha^{(1)})}{P_s(x_{ce})}, \dots, \frac{P_s(x_{ce} | \alpha^{(N_{sim})})}{P_s(x_{ce})} \right]^T$$



Parameter point for inference  
 $\alpha$

Predictions

$$p(\Delta'_{scb} | \alpha, A_s^{sim}, \Delta_{scb}(A_s^{sim})) = \mathcal{N}(\bar{\Delta}'_{scb}, \text{COV}(\Delta'_{scb}))$$



The final **posterior** prediction is a probability distribution of values

normal distribution with GP predicted mean and covariance kernel

# Why HistFactory v2 for SBI

- Even when the systematics model is built so that each element in the high-dimensional parameter space is independent of the others, there is no guarantee that the impact on the final likelihood factorizes.
- Theoretical systematic models are often chosen ad-hoc without strong guarantee that each parameter is even independent of others.
- **With Simulation-Based Inference, we are probing the entire high-dimensional parameter space (POIs and NPs) using high-dimensional kinematics, significantly improving the precision for all parameters.**
- More important than ever to build a robust model across the parameter space. More details in the talk tomorrow.