ICCUB Winter Meeting 2025

# Improving Statistical Analyses in Particle Physics Phenomenology

B. Capdevila University of Barcelona, ICCUB

Institut de Ciències del Cosmos
UNIVERSITAT DE BARCELONA
EXCELENCIA MARÍA DE MAEZTU 2020–2023

"la Caixa" Foundation

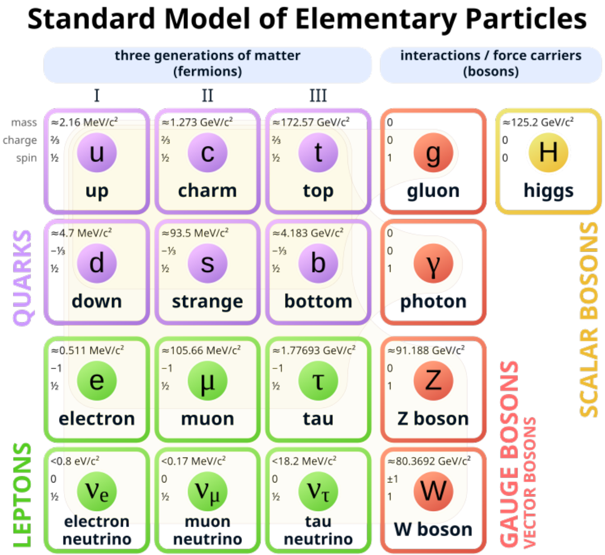
European Commission

*In collaboration with:*

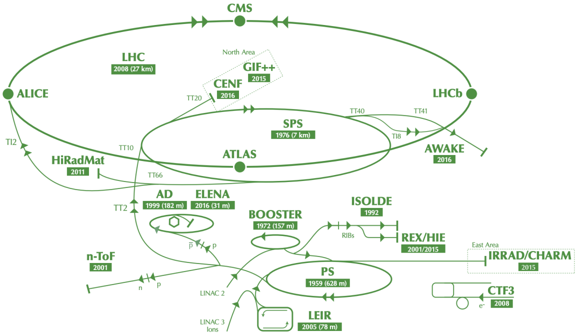
**A. Scaffidi; arXiv:250x.xxxx**

# Outline

# Basic Ideas in Particle Physics Phenomenology


Standard Model of Elementary Particles
three generations of matter (fermions)
interactions / force carriers (bosons)
I
II
III
mass
charge
spin
≈2.16 MeV/c²
⅔
½
u
up
≈1.273 GeV/c²
⅔
½
c
charm
≈172.57 GeV/c²
⅔
½
t
top
0
0
1
g
gluon
≈125.2 GeV/c²
0
0
H
higgs
≈4.7 MeV/c²
−⅓
½
d
down
≈93.5 MeV/c²
−⅓
½
s
strange
≈4.183 GeV/c²
−⅓
½
b
bottom
0
0
1
γ
photon
≈0.511 MeV/c²
−1
½
e
electron
≈105.66 MeV/c²
−1
½
μ
muon
≈1.77693 GeV/c²
−1
½
τ
tau
≈91.188 GeV/c²
0
1
Z
Z boson
<0.8 eV/c²
0
½
νe
electron neutrino
<0.17 MeV/c²
0
½
νμ
muon neutrino
<18.2 MeV/c²
0
½
ντ
tau neutrino
≈80.3692 GeV/c²
±1
1
W
W boson
QUARKS
LEPTONS
GAUGE BOSONS
VECTOR BOSONS
SCALAR BOSONS

# The LHC Era of Particle Physics



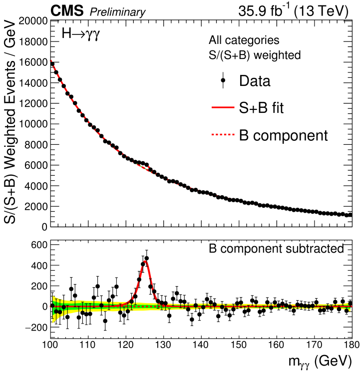

⇒ We explore this theory (and potentially others) using **Particle Colliders**

⇒ The current era of Particle Physics is shaped significantly by the LHC, the world's **largest** and **most powerful** collider ever built

⇒ Several other colliders and accelerators are operational, fulfilling more specialised roles

- ⇒ SuperKEKB (Belle / Belle II)
- ⇒ The Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory
- ⇒ Many others, including dedicated accelerator experiments, beam experiments, . . .

# The Higgs Boson Discovery




CMS, CMS-PAS-HIG-17-015

⇒ The discovery of the **Higgs boson** in 2012 stands as the crowning achievement of the LHC era

⇒ We usually say that *this discovery marked the completion of the Standard Model*, but did it?

- ⇒ The precise nature of the Higgs boson remains an open question
- ⇒ With a large portion of the Higgs potential parameter space still unexplored

# The Current Paradigm

$$\begin{aligned}\mathcal{L}_{\rm SM} = i\left(\bar{\ell}_L \not{D} \ell_L + \bar{e}_R \not{D} e_R + \overline{Q}_L \not{D} Q_L + \bar{u}_R \not{D} u_R + \bar{d}_R \not{D} d_R\right) \\ - \frac{1}{4} G^A_{\mu\nu} G^{A\mu\nu} - \frac{1}{4} W^I_{\mu\nu} W^{I\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \\ + \left(D_\mu \phi\right)^\dagger \left(D^\mu \phi\right) - \mu^2 \phi^\dagger \phi - \lambda \left(\phi^\dagger \phi\right)^2 \\ - \left(\bar{\ell}_L Y^\ell \phi e_R + \overline{Q}_L Y^u \epsilon \phi^* u_R + \overline{Q}_L Y^d \phi d_R + \text{h.c.}\right)\end{aligned}$$

- ⇒ The Standard Model is arguably the *most successful* theory of fundamental physics
- ⇒ **However**, it is clear that it remains incomplete ...
  - ⇒ **Gravity**, **Dark Matter**, and **Dark Energy** are not accounted for within the SM

... and it presents several unresolved questions, often referred to as *problems* or *puzzles*:

- ⇒ *Hierarchy Problem*: stabilising the Higgs boson requires **large cancellations** of **quantum corrections**
- ⇒ *Strong CP Problem*: the CP-violating term $\mathcal{L}_{SM} \supset \theta \frac{g_s^2}{32\pi^2} G_{\mu\nu} \tilde{G}^{\mu\nu}$ allowed, yet experimentally $\theta \sim 0$
- ⇒ *Flavour Puzzle*: why are there **three families** of **quarks** and **leptons**?
- ⇒ *Neutrino masses*: not incorporated in the SM
- ⇒ *Baryon Asymmetry Problem*: the CP-violation present in $\mathcal{L}_{\rm SM}$ is insufficient to explain the observed matter-antimatter asymmetry
- ⇒ ...

# Towards a Deeper Understanding of the SM and Beyond

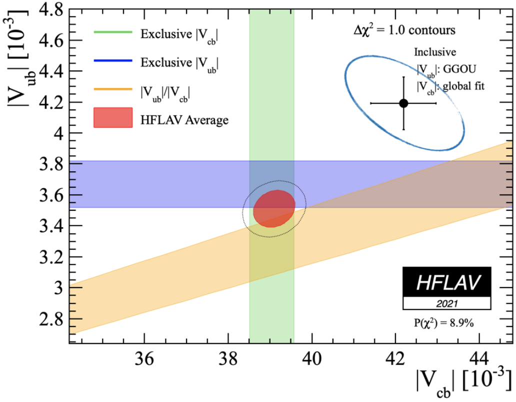
$$
\begin{aligned}
\mathcal{L}_{\text{BSM}} \supset\, & i\left(\bar{\ell}_L \not{D} \ell_L + \bar{e}_R \not{D} e_R + \overline{Q}_L \not{D} Q_L + \bar{u}_R \not{D} u_R + \bar{d}_R \not{D} d_R\right) \\
& - \frac{1}{4} G^A_{\mu\nu} G^{A\mu\nu} - \frac{1}{4} W^I_{\mu\nu} W^{I\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \\
& + \left(D_\mu \phi\right)^\dagger \left(D^\mu \phi\right) - \mu^2 \phi^\dagger \phi - \lambda \left(\phi^\dagger \phi\right)^2 \\
& - \left(\bar{\ell}_L Y^\ell \phi e_R + \overline{Q}_L Y^u \epsilon \phi^* u_R + \overline{Q}_L Y^d \phi d_R + \text{h.c.}\right) + \ldots
\end{aligned}
$$

⇒ Despite the many **indications of potential shortcomings** in the SM, no **conclusive evidence** against it has been found in current data

⇒ The unresolved puzzles and missing elements necessitate **further investigation**

- ⇒ **Mathematical approach**: Quantum Gravity, extended symmetries (model building), formal QFT, . . .
- ⇒ **Phenomenological approach**: understanding how data aligns with the SM (**anomaly detection**), refining SM predictions, exploring extended symmetries (model building), determining SM fundamental parameters, . . .

# Determinations of $V_{cb}$ and $V_{ub}$



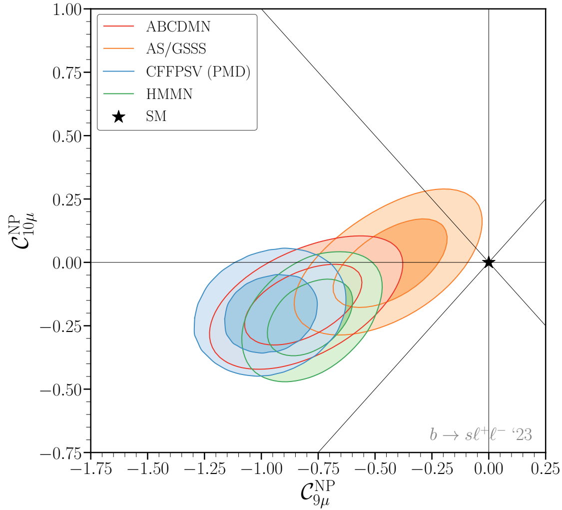

BC, Gambino, Nandi; **JHEP 04 (2021) 137** [arXiv:2102.03343]
Bordone, BC, Gambino; **Phys. Lett. B 822 (2021) 136679** [arXiv:2107.00604]

⇒ $V_{ub}$ and $V_{cb}$ are two elements of the **CKM matrix** that governs *quark-level transitions*

⇒ These elements arise from the diagonalisation of the *quark mass matrix*

$$u_L^i \to U_u^{ij} u_L^j, \quad d_L^i \to U_d^{ij} d_L^j, \quad V_{\text{CKM}} = U_u^\dagger U_d$$

# Anomaly Detection and Characterisation



# Anomaly Detection

# Anomalies in Particle Physics

⇒ The term **anomaly** has different meanings in the context of Particle Physics

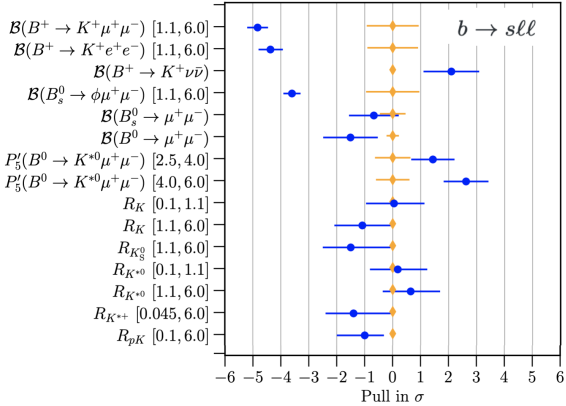
⇒ In QFT, an anomaly is a *symmetry* of the *classical theory* that is *broken* by *quantum corrections*

⇒ In the context of *phenomenology*, anomalies refer to **deviations** between **experimental data** and **theoretical predictions** under a given **null hypothesis**

⇒ **Goal:** to *properly* quantify the **statistical significance** of observed anomalies

# Recent Anomalies in Particle Physics

- ⇒ $B$-meson[1] anomalies ($b \to s\ell^+\ell^-$, $b \to c\ell\nu$)
- ⇒ $(g-2)_\mu$
- ⇒ $V_{cb}$, $V_{ub}$ puzzle
- ⇒ ...




[1] A $B$-meson is a $b$-quark and a light-quark bound state, i.e. $B^0$ ($\bar{b}d$), $B^+$ ($\bar{b}u$), $B_s$ ($\bar{b}s$), $B_c$ ($\bar{b}c$), ...

# Traditional Frequentist Analyses

⇒ In **conventional analyses**, Gaussian likelihoods are assumed

$$p(\boldsymbol{x}|H_i) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_{H_i}, \boldsymbol{\Lambda}_{H_i}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Lambda}_{H_i}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_{H_i})^T \boldsymbol{\Lambda}_{H_i}^{-1} (\boldsymbol{x} - \boldsymbol{x}_{H_i})\right)$$

- ⇒ $\boldsymbol{\Lambda}_{H_i}$ is the **sum of theoretical and experimental covariances**

$$\boldsymbol{\Lambda}_{H_i} = \boldsymbol{\Lambda}_{H_i}^{\text{th}} + \boldsymbol{\Lambda}^{\text{exp}}$$

⇒ Measuring goodness-of-fit

- ⇒ Use $-\log p(\boldsymbol{x}|H_i)$ as a **statistic** to measure **agreement between data and hypothesis** $H_i$
- ⇒ If **theoretical predictions** and **experimental data** are **normally distributed**, $-\log p(\boldsymbol{x}|H_i)$ follows a $\chi^2$-distribution with $n_{\text{dof}} = n_{\text{obs}}$ in the analysis
- ⇒ If **not normally distributed**, $-\log p(\boldsymbol{x}|H_i)$ is **only asymptotically** $\chi^2$ due to the central limit theorem
- ⇒ Assuming $\chi^2$-distribution when it is not **creates biases** in calculating $p$-values from $-\log p(\boldsymbol{x}_{\text{exp}}|H_i)$

# Experimental Data in Phenomenological Analyses

- ⇒ Experimental data:
  - ⇒ Released from the experiments as a **vector of means** $\mu^{\text{exp}}$ and a **covariance matrix** $\Lambda^{\text{exp}}$
- ⇒ **Implicitly assumes a Gaussian distribution** for the experimental measurements

$$p(\boldsymbol{x}^{\text{exp}}) = \mathcal{N}(\boldsymbol{x}^{\text{exp}}; \boldsymbol{\mu}^{\text{exp}}, \boldsymbol{\Lambda}^{\text{exp}})$$

| | $4.0 < q^2 < 6.0\,\text{GeV}^2/c^4$ |
|---|---|
| $P_1$ | $0.088 \pm 0.235 \pm 0.029$ |
| $P_2$ | $0.105 \pm 0.068 \pm 0.009$ |
| $P_3$ | $-0.090 \pm 0.139 \pm 0.006$ |
| $P_4'$ | $-0.312 \pm 0.115 \pm 0.013$ |
| $P_5'$ | $-0.439 \pm 0.111 \pm 0.036$ |
| $P_6'$ | $-0.293 \pm 0.117 \pm 0.004$ |
| $P_8'$ | $0.166 \pm 0.127 \pm 0.004$ |

| | $F_L$ | $P_1$ | $P_2$ | $P_3$ | $P_4'$ | $P_5'$ | $P_6'$ | $P_8'$ |
|---|---|---|---|---|---|---|---|---|
| $F_L$ | 1.00 | 0.04 | 0.05 | −0.10 | −0.04 | −0.14 | −0.17 | 0.14 |
| $P_1$ | | 1.00 | 0.06 | 0.07 | −0.06 | −0.10 | −0.03 | 0.02 |
| $P_2$ | | | 1.00 | −0.02 | −0.14 | −0.09 | −0.03 | −0.01 |
| $P_3$ | | | | 1.00 | −0.01 | 0.07 | 0.19 | −0.01 |
| $P_4'$ | | | | | 1.00 | 0.02 | 0.04 | 0.01 |
| $P_5'$ | | | | | | 1.00 | 0.09 | 0.00 |
| $P_6'$ | | | | | | | 1.00 | 0.02 |
| $P_8'$ | | | | | | | | 1.00 |

LHCb, arXiv:2003.04831

# Theoretical Predictions in Phenomenological Analyses

⇒ Theoretical predictions

- ⇒ For the observables in the analysis $\boldsymbol{x} = (x_1, \ldots, x_n)$, we have functions representing their **theoretical predictions** in terms of several **input parameters** $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m)$

$$x_i = x_i(\nu_1, \ldots, \nu_m)$$

- ⇒ The **input parameters** are distributed according to some distribution, usually Gaussian

$$\boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{\nu}; \boldsymbol{\mu}_\nu, \boldsymbol{\Lambda}_\nu)$$

where $\boldsymbol{\mu}_\nu$ and $\boldsymbol{\Lambda}_\nu$ means and covariance of the distribution of underlying parameters

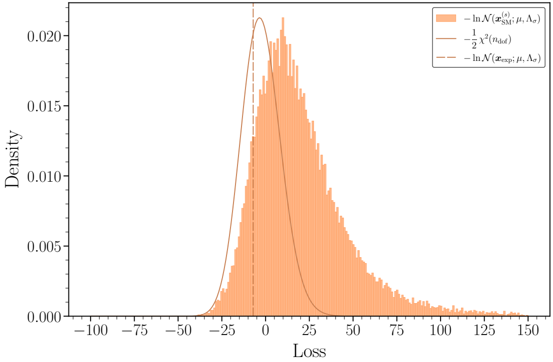
⇒ Implications

- ⇒ Even if the distribution of parameters is Gaussian, observables with complex structures **do not distribute normally**
- ⇒ **Only** observables with a **linear dependence** on the underlying parameters
- ⇒ The **likelihood** $p(\boldsymbol{x}|H_i)$ will generally be distributed under a **non-Gaussian distribution**

# Distribution of $-\log p(\boldsymbol{x}|H_i)$ vs Asymptotic $\chi^2$

⇒ Using samples $x_{\text{SM}}^{(s)}$ of SM predictions for the $b \to s\ell\ell$ observables dataset, we calculated $-\log p(\boldsymbol{x}|H_0)$ for each sample ⇒ log-likelihood distribution

⇒ Sizeable difference between the $-\log p(\boldsymbol{x}|H_i)$ and the asymptotic $\chi^2$ distribution with corresponding degrees of freedom



# Estimating Likelihoods with Full Generality

- ⇒ Goal: determine the full likelihood $p(\boldsymbol{x}|H_i) = p(\boldsymbol{x})$ **without introducing bias**
- ⇒ Known: distribution of underlying inputs $p(\boldsymbol{\nu})$ (model parameters)
- ⇒ Computable: $p(\boldsymbol{x}|\boldsymbol{\nu})$, achievable by simulating observables $\boldsymbol{x}$ using sampled parameters $\boldsymbol{\nu}$ from their known distributions
- ⇒ Explicitly calculate $p(\boldsymbol{x})$ as a marginal likelihood

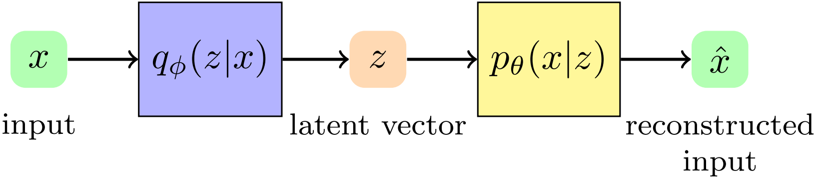
$$p(\boldsymbol{x}) = \int d\boldsymbol{\nu}\; p(\boldsymbol{x}|\boldsymbol{\nu})p(\boldsymbol{\nu})$$

  - ⇒ The **curse of dimensionality**: in most real-life applications $\boldsymbol{\nu}$ is usually high-dimensional
  - ⇒ Challenge: a direct computation the $d\boldsymbol{\nu}$ integral is generally **intractable** or **computationally very expensive**

- ⇒ Using **Machine Learning** techniques to learn the full likelihood function
  - ⇒ Using **Normalising Flows** to encode the densities $p(x_i|H_0)$
  - ⇒ **Variational Autoencoders** (VAEs) provide a feasible approach to approximate $p(\boldsymbol{x})$ with **arbitrary precision**

# Anomaly Detection with VAEs

# Introducing Variational Autoencoders




- ⇒ VAE framework
  - ⇒ **Model**: Pairs a probabilistic decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ with a probabilistic encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x})$
  - ⇒ **Latent variable $\boldsymbol{z}$** approximates underlying parameters $\boldsymbol{\nu}$
  - ⇒ VAEs **do not** map inputs to a deterministic latent variable, but to a **probability space** $p(\boldsymbol{z})$
  - ⇒ $\theta$: parameters of the decoder
  - ⇒ $\phi$: parameters of the decoder

# The Variational Lower Bound

- ⇒ The Variational Lower Bound (ELBO) relates to two joint probability density functions: $p_\theta$ and $q_\phi$

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]$$

  - ⇒ $p_\theta(\boldsymbol{x}, \boldsymbol{z})$: joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$
  - ⇒ $q_\phi(\boldsymbol{z}|\boldsymbol{x})$: approximate encoder posterior
  - ⇒ Simplifies to

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$$

- ⇒ Includes Kullback-Leibler divergence (KL-div) which is a distance in distribution space

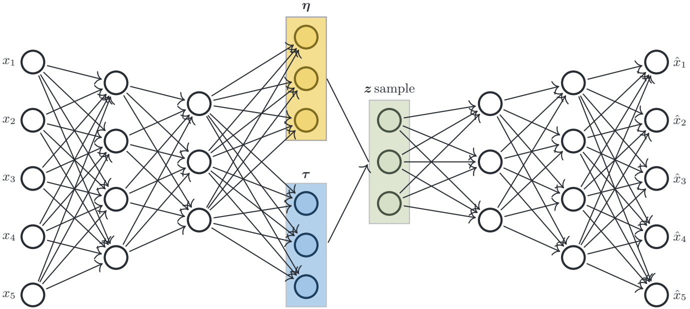
$$D_{KL}(q_\phi(w)||p_\theta(w)) = \mathbb{E}_{q_\phi(w)}\left[\log \frac{q_\phi(w)}{p_\theta(w)}\right]$$

- ⇒ ELBO and VAE objective function
  - ⇒ Log-likelihood relation

$$\log p_\theta(\boldsymbol{x}) \geq \mathcal{L}(\theta, \phi; \boldsymbol{x})$$

  - ⇒ Maximize ELBO to approximate the true log-likelihood

# Deep Learning Implementation of Variational Autoencoders




⇒ Implementation and parametrisation

- ⇒ **Neural Networks as parametrisers**
- ⇒ $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ parameterised using deep neural networks

$$\theta = \{W_{l_1}, \ldots, W_{l_L}, b_{l_1}, \ldots, b_{l_L}\}$$
$$\phi = \{V_{l_1}, \ldots, V_{l_L}, c_{l_1}, \ldots, c_{l_L}\}$$

where $W_l$ $(V_l)$, $b_l$ $(c_l)$ are the weights and biases of the encoder (decoder) network

- ⇒ This setup enables the modeling of complex, non-linear relationships between observed data and latent variables

# Training the VAE with Theoretical and Experimental Inputs

⇒ Generating Theoretical Predictions

- ⇒ Start by sampling the distribution of underlying inputs under hypothesis $H_0$, $p(\boldsymbol{\nu}|H_0)$
- ⇒ Compute the vector of observables $\boldsymbol{x}^s$ for these values to obtain a sample of theoretical predictions: $\boldsymbol{x}^s = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{n_{\text{sample}}})$

⇒ Incorporating Experimental Uncertainties

- ⇒ Smear the samples with experimental uncertainties to simulate realistic observational data

$$\boldsymbol{x}'^s = \boldsymbol{x}^s + \boldsymbol{L}_{\boldsymbol{\Lambda}^{\text{exp}}} \boldsymbol{w}$$

- ⇒ $\boldsymbol{L}_{\boldsymbol{\Lambda}^{\text{exp}}}$: Cholesky decomposition of the experimental covariance matrix $\boldsymbol{\Lambda}^{\text{exp}}$
- ⇒ $\boldsymbol{w} \sim \mathcal{N}(w; \mathbf{0}, \mathbf{1})$: Normal noise vector simulating experimental noise

⇒ Training the VAE

- ⇒ Divide the smeared dataset into **training and testing datasets**
- ⇒ Use the training dataset to **optimise the parameters** of the VAE, **minimising** the -ELBO

$$\mathcal{L}(\theta, \phi; \boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \beta D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) || p_\theta(\boldsymbol{z}))$$

- ⇒ Objective: **approximate the full log-likelihood distribution** of the observables under $H_0$ as closely as possible

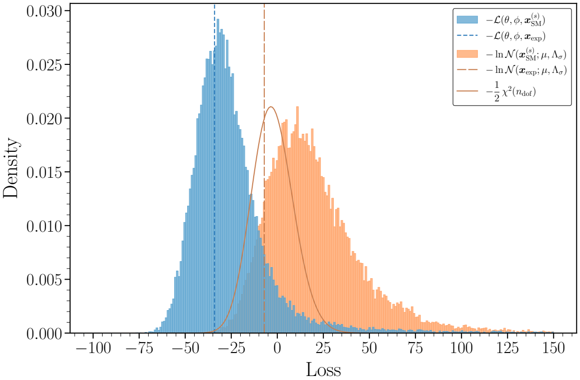
# Statistical Analysis Using Trained VAE for $b \to s\ell^+\ell^-$

⇒ Analysing the test dataset

⇒ Compute the -ELBO distribution using the test dataset, approximating the full -log-likelihood under hypothesis $H_0$.

⇒ Evaluate -ELBO for the experimental data to compute the $p$-value

⇒ Preliminary results for the $b \to s\ell\ell$ dataset



# Outlook

# Outlook and Continued Research

- ⇒ Deepening understanding of VAE parameters
  - ⇒ Exploring how different parameters influence the anomaly score and VAE's generative properties
- ⇒ Ongoing hyperparameter optimisation
  - ⇒ Continuously refining the model to enhance its predictive accuracy and anomaly detection
- ⇒ Addressing sparse covariance matrices
  - ⇒ Using random matrix theory techniques to sample observables and covariance matrices at the same time (LKJ distribution, Wishart distribution)
  - ⇒ Will allow us to quantify the uncertainty attached to many unnatural zeros in the experimental covariance matrix
- ⇒ Expanding application scope
  - ⇒ Applying methodologies to SMEFT fits beyond $b \to s\ell\ell$

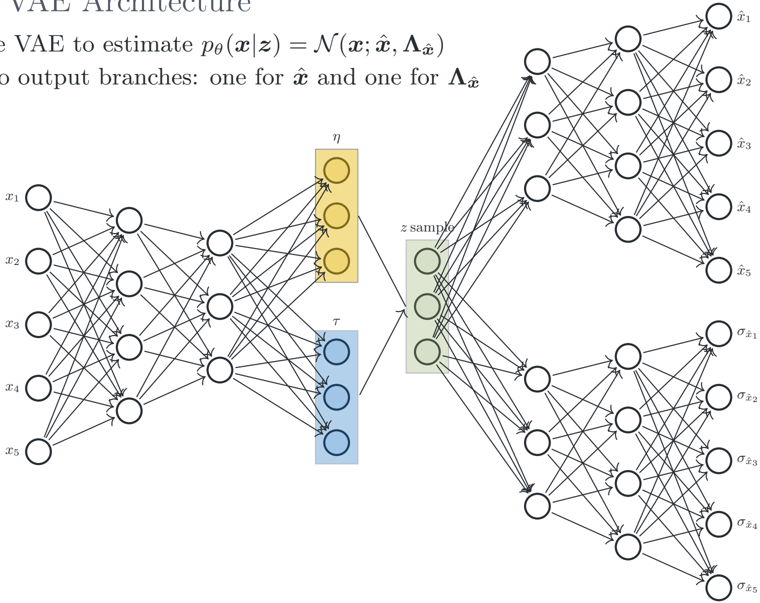
Thank You!

# Backup Slides

# Dual-Branch VAE Architecture

⇒ We need the VAE to estimate $p_\theta(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \hat{\boldsymbol{x}}, \boldsymbol{\Lambda}_{\hat{\boldsymbol{x}}})$

⇒ We need two output branches: one for $\hat{\boldsymbol{x}}$ and one for $\boldsymbol{\Lambda}_{\hat{\boldsymbol{x}}}$



# Refining Model Tuning and Validity in VAE Training

- Challenges in Model Tuning
  - ⇒ How do we determine the optimal dimensionality for the DNNs of the encoder and decoder, or the correct value of $\beta$?
  - ⇒ Could these choices bias the $p$-value?
  - ⇒ The choice of neural networks' architecture and $\beta$ significantly affects the model's performance and the fidelity of the statistical results
- Testing and Validating Model Parameters
  - ⇒ Ongoing research and empirical testing are essential to optimise these parameters while minimising biases
- Strategies for Validation and Hyperparameter Optimisation
  - ⇒ Employ validation techniques to ensure model outputs are stable and reliable across various parameter configurations
  - ⇒ Use synthetic datasets to evaluate the impact of hyperparameter adjustments on model performance

# Optimising Hyperparameters

- Generating artificially anomalous data
  - ⇒ Used to train the VAE across different configurations to optimise anomaly detection
- Tuning $\beta$ in the ELBO
  - ⇒ Exploring the impact of $\beta$ on anomaly detection and VAE's generative accuracy
- Pre-experimental blind analysis
  - ⇒ Ensures that the final measurement of experimental data is unbiased