# Search for new physics through Higgs-pair production with the ATLAS experiment at the LHC

**Thomas Ghorbanian**

Supervised by
Prof. Nikolaos Konstantinidis

Deptartment of Physics & Astronomy
University College London

## Acknowledgements

My completion of this project could not have been accomplished without the support and guidance of Prof. Nikolaos Konstantinidis, Lucas Borgna and the UCL $hh \to 4b$ team. I offer my sincere appreciation for the learning opportunities provided by the team.

I also thank Nicole Hartman from the SLAC National Accelerator Laboratory at Stanford. The exceptional ideas she shared with me formed the foundation of this project.

# Contents

# 1 Introduction

Particle physics studies the nature of the fundamental constituents of matter and radiation. Via an elegant set of equations called the Standard Model (SM), our current attempt at describing everything we observe in the universe is underpinned by our understanding of quantum fields and their interactions. With the discovery of the Higgs boson at the Large Hadron Collider (LHC) [1], all of the particles in the Standard Model have now been observed, demonstrating a remarkable success of modern physics.

However, while exceptionally accurate, the Standard Model is known to be incomplete. It fails to account for some observed phenomena, such as gravitational interactions, Dark Matter, the origin of neutrino mass and oscillations, and matter-antimatter asymmetry. Through the investigation of Higgs-pair production at the LHC, this project aims to search for potential physics Beyond the Standard Model (BSM). In particular, with the rapidly developing landscape of deep learning, a Bayesian approach to background estimation is evaluated as a tool for modelling $hh \to 4b$ decay.

## 1.1 The Standard Model

Developed during the second half of the $20^{th}$ century, the Standard Model is a Quantum Field Theory (QFT) describing the mechanisms underlying the strong nuclear, weak nuclear and electromagnetic interactions [2–4]. The interactions emerge from local symmetries of the universe; physics invariant under the the $SU(3)_C$ gauge group describe the strong sector (quantum chromodynamics), while physics invariant under the $SU(2)_L \otimes U(1)_Y$ gauge group describe the electroweak sector (quantum electrodynamics and the weak force).

As a QFT, particles are treated as excitations of quantum fields. The Standard Model deals with three types of field; vector fields (describing interactions), spinor fields (describing 'matter') and scalar fields (describing the electroweak breaking sector). The equations of motions of free fields as well as their interactions are encoded in the Standard Model Lagrangian:

$$\mathcal{L}_{SM} = -\frac{1}{4}(F^{\alpha}_{\mu\nu})^2 + i\bar{\psi}\gamma^{\mu}D_{\mu}\psi + y\bar{\psi}\psi\phi + \mid D_{\mu}\phi \mid^2 - \mu^2\phi^{\dagger}\phi - \lambda(\phi^{\dagger}\phi)^2, \tag{1}$$

where $F^{\alpha}_{\mu\nu} = \partial_{\mu}A^{\alpha}_{\nu} - \partial_{\nu}A^{\alpha}_{\mu} + gf^{abc}A^{b}_{\mu}A^{c}_{\nu}$ is the field strength tensor and $D_{\mu} = \partial_{\mu} - igA^{\alpha}_{\mu}t^{\alpha}$ is the covariant derivative, with gauge fields $A$, spinor fields $\psi$ and the Lorentz-scalar Higgs field $\phi$. $f^{abc}$ represent the fine structure constants, $t^{\alpha}$ are the corresponding gauge group representation matricies, and $\gamma^{\mu}$ are the Dirac matricies. The first term $-\frac{1}{4}(F^{\alpha}_{\mu\nu})^2$ captures the dynamics of gauge fields that give rise to the force-mediating gauge bosons: the gluon, the photon, and the weak bosons. The second term defines the kinematics of fermions and their interaction with gauge bosons. The third term describes the interactions of the Higgs field with fermions through the Yukawa couplings, $y$. The fourth term describes the dynamics of the Higgs field and its interaction with gauge fields. The fifth term describe the self-interactions of the Higgs field.

Of particular interest are the final three terms describing the Higgs sector. Following the discovery of the Higgs boson at the LHC, ATLAS and CMS have been studying the coupling of the Higgs to the SM as a means of probing BSM physics [5].

## 1.2 The Higgs Boson

Equation (1) suggests that the Higgs field is the only massive field, since explicit fermion and gauge boson mass terms in the Lagrangian would break local symmetry and lead to an unphysical theory. However, by setting:

$$\phi(x) = \frac{1}{\sqrt{2}}(v + h(x)), \tag{2}$$

and expanding the field about the minima, $v$, of the potential described by the last two terms of (1):

$$V(\phi) = \mu^2\phi^{\dagger}\phi + \lambda(\phi^{\dagger}\phi)^2, \tag{3}$$

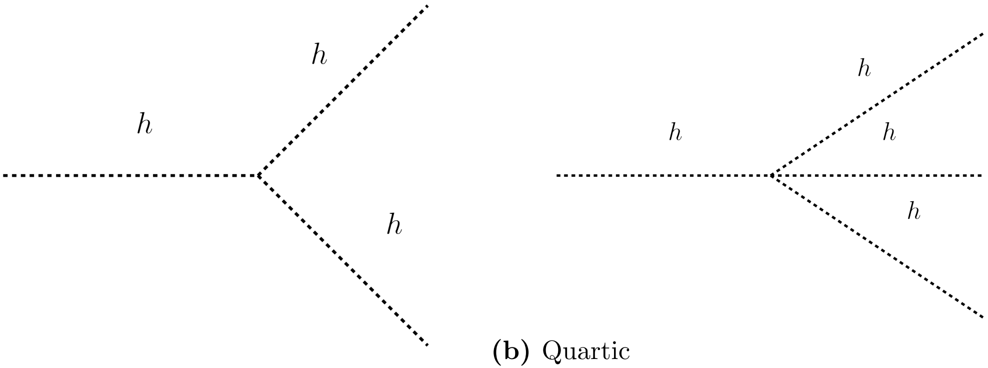
rotational $U(1)$ symmetry is *spontaneously broken*, giving rise to a Higgs field $h(x)$ and a Higgs boson of mass $m_h = \sqrt{2\lambda}v$. $h(x)$ is an essential component of the SM responsible for generating the masses of quarks and charged leptons (through Yukawa couplings [3]) and the electroweak gauge bosons (through the Higgs mechanism [6, 7]), *without* introducing explicit mass terms in the Lagrangian.

Furthermore, the expansion about $v$ adds interaction terms of the form:

$$\mathcal{L} \propto -\lambda v h^3 - \frac{1}{4}\lambda h^4, \tag{4}$$

to the Lagrangian, describing the Higgs boson self-interaction via *trilinear self-coupling* $\lambda_{hhh} = \lambda v$ and *quartic coupling* $\lambda_{hhhh} = \frac{\lambda_{hhh}}{4v}$, shown in Figure 1. The SM thus predicts multi-$h$ *production* with well defined cross-sections.

Probing the shape of the Higgs potential can provide insights on the electroweak phase transition during the early universe [8, 9] and in turn shed light on matter-antimatter asymmetry [10]. With $\mu$ already measured by Higgs searches, a measurement of $\lambda_{hhh}$ would therefore be a well motivated area of research in collider physics, providing an independent test of the global properties of the Higgs potential. In addition, it would present an opportunity to test the SM Higgs self-interaction cross-section against the predictions of BSM theories.



**(a)** Trilinear

**(b)** Quartic

**Figure 1** Diagram depicting Higgs boson trilinear and quartic self-couplings.

## 1.3 Di-Higgs Production

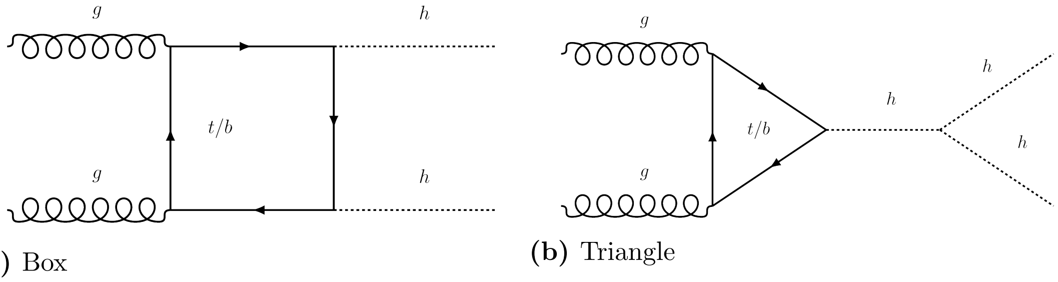
At the LHC, tri-linear coupling can be directly probed by measuring the cross section of Higgs boson pair-production (*di-Higgs* production). While well defined, the SM predicts the cross section of di-Higgs production to be small ($\sigma_{hh} \sim 34$ fb) [11]. Since many BSM theories predict enhanced rates of di-Higgs production, any experimental deviation from this small production rate could be an indication of BSM contributions. For *non-resonant* production in particular (Figure 4a), the SM coupling ratio:

$$\kappa_\lambda = \frac{\lambda_{hhh}}{\lambda_{hhh}^{SM}} \tag{5}$$

is investigated.

### 1.3.1 Di-Higgs Production in the Standard Model

The SM predicts gluon-gluon fusion to be the main production channel for di-Higgs production. The small cross section is a result of the interaction occurring in a reduced phase space (where two heavy particles are in the final state) and the leading order diagrams, shown in Figure 2, undergoing destructive interference. The resulting cross section of various Higgs boson production modes as a function of $\kappa_\lambda$ is shown in Figure 3.



**(a)** Box

**(b)** Triangle

**Figure 2** Leading order di-Higgs production modes via gluon-gluon fusion.

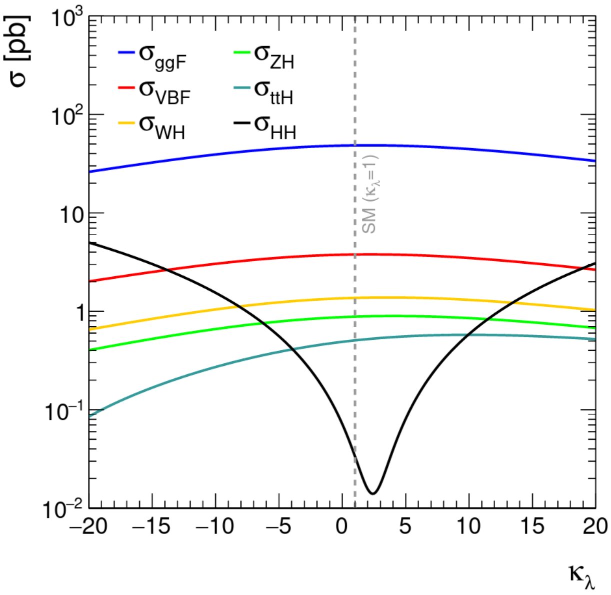
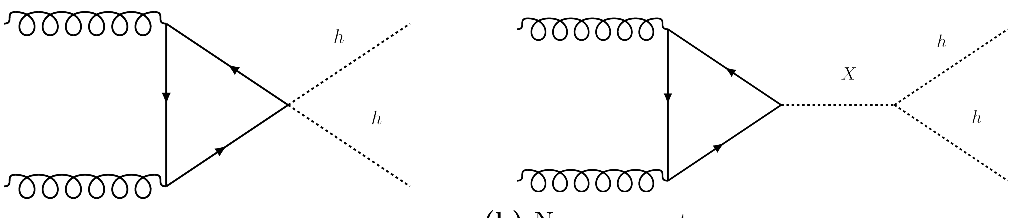


**Figure 3** Single-Higgs and double-Higgs production cross section as a function of $\kappa_\lambda$ [12]. The dashed line intercepts the values corresponding to the SM hypothesis ($\kappa_\lambda = 1$).

### 1.3.2 Di-Higgs Production Beyond the Standard Model

Many BSM theories predict enhanced rates of either resonant or non-resonant di-Higgs production (see Figure 4). Non-resonant di-Higgs enhancement modifies $\lambda_{hhh}$ or activates new vertices, resulting to non-standard couplings or new heavy states running in loops. Resonant di-Higgs enhancement predict heavy resonances decaying to a pair of Higgs bosons.



**(a)** Resonant

**(b)** Non-resonant

**Figure 4** An example of resonant and non-resonant di-Higgs production. Heavy resonances depicted by $X$.

Two-Higgs-doublet models (2HDM) [13, 14] are amongst the simplest extensions to the SM, containing two Higgs isospin doublets instead of just one. The addition of the second Higgs doublet would lead to a richer phenomenology of scalar states after spontaneous symmetry breaking, predicting *five* physical Higgs bosons. The heavier scalars would couple to the $m_h = 125\,\mathrm{GeV}$ boson, and decay to a $hh$ pair if kinematically allowed. In addition to possibly solving some of the unanswered questions of the SM, such as baryogenesis or dark matter, two-Higgs-doublet models can be integrated into more complex models that stand as possible candidates for a fundamental theory, such as supersymmetry.

Randall-Sundrum (RS) models [15] offer an integration of gravity into the SM by adding a (compact) 5th dimension that exists in a separate brane from our usual 4-dimensional (3 space + 1 time) brane. SM particles would be localised on the 4-dimensional brane, and gravitons would propagate along the 5th dimension. The model explains naturally why the gravitational force is so small, since gravitons would travel mostly in dimensions we do not observe. Kaluza-Klein modes of the graviton would appear on our 4-dimensional brane as massive particles. If such excitations exist, they would couple to the Higgs boson and could be observed as a resonance in the di-Higgs mass spectrum.

The analysis described in this report focuses on non-resonant signal models. For example, composite Higgs models (CHM) [16, 17] where the Higgs boson is a bound state of new strong interactions, or models with light coloured scalars [18], are expected to increase production of non-resonant Higgs boson pairs. Since the different spin of the resonances leads to different kinematics of di-Higgs decay, resonant and non-resonant searches require slightly different sensitivities.

### 1.4 The $hh \to 4b$ Search

Following the discovery of the Higgs boson at the LHC, ATLAS and CMS have investigated various di-Higgs decay channels, notably $bb\bar{b}\bar{b}$ [19, 20], $b\bar{b}W^+W^-$ [21], $b\bar{b}\tau^+\tau^-$ [22], $W^+W^-W^+W^-$ [23], $b\bar{b}\gamma\gamma$ [24] and $W^+W^-\gamma\gamma$ [25], using data from both Run I ($s = \sqrt{8}$ TeV) and Run II ($s = \sqrt{13}$ TeV) of the LHC.

A decay mode of particular interest is $bb\bar{b}\bar{b}$ ($hh \to 4b$), with a final state containing two $b$ quark-antiquark pairs. Out of the phenomenologically rich set of final states, $hh \to 4b$ is by far the most dominant decay mode, as shown in Figure 5. Its search can be conducted in one of two regimes; either in the region of phase space where one pair is highly Lorentz-boosted and is reconstructed as a single large-area jet (boosted regime), or the pair is resolved and reconstructed using two small-radius jets (resolved regime), as shown in Figure 6. Multivariate analysis can then place an upper limit on the cross section of di-Higgs production, allowing for a direct comparison with SM and BSM predictions. Existing studies thus far [19, 20] have yet to report significant data excess above the estimated background of SM predictions.

However, a significant challenge in the search for $hh \to 4b$ is the extraction of signals amongst a large multi-jet background. Current studies employ data-driven multivariate analysis, such as Boosted Decision Trees (BDT) and Neural Networks (NN), to reweight multi-jet events with 4 jets, two of which are $b$-tagged (so-called $2b$ events hereafter), to make them kinematically as similar as possible to multi-jet events with four $b$-tagged jets (referred to $4b$ events hereafter). The hypothesis is that the kinematics of multiple $b$-tagged jets are similar to the kinematics of events with four $b$-tagged jets, and any differences can be corrected by a reweighting procedure. Thus, the foundation of existing background estimates lie in the derivation of a reweighting function:

$$w(\mathbf{x}) = \frac{p_{4b}(\mathbf{x})}{p_{2b}(\mathbf{x})}, \tag{6}$$

where $p_{4b}(\mathbf{x})$ and $p_{2b}(\mathbf{x})$ are the probability density functions for $4b$ and $2b$ events respectively across some kinematic variables $\mathbf{x}$. In this scheme, the probability density functions are defined by histograms.

This project assumes an *alternative* strategy for $4b$ background estimation that does not require a reweighting procedure. No assumptions are made about the kinematic similarity of $4b$ events with other multiple $b$-tagged events. Instead, an *expressive* probability density function for $4b$ events is learnt *directly* in a Bayesian manner via probabilistic deep learning. In addition to background estimation, the scheme allows for statistical inference and can be used with any general purpose sampling algorithm. The motivation of such a procedure is to build a full, expressive statistical description of $b$-tagged events with a minimal number of assumptions.

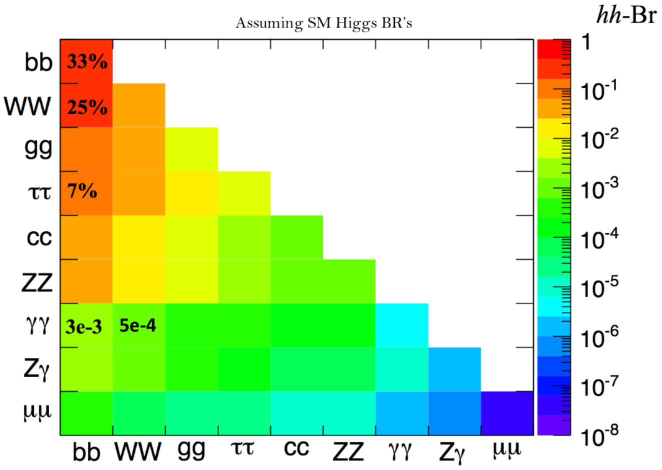
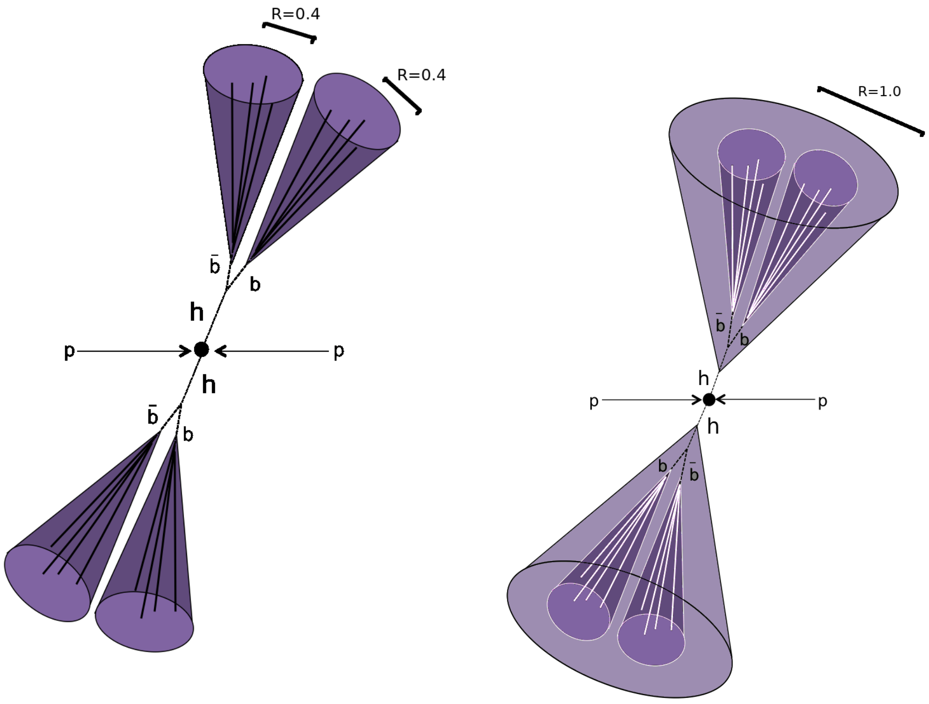


**Figure 5** Di-Higgs branching ratios at the LHC. $hh \to 4b$ is the most common decay mode.



**(a)** Resolved

**(b)** Boosted

**Figure 6** Search for $hh \to 4b$ via (a) resolved analysis, requiring 4b-tagged small-R jets ($R = 0.4$) (b) boosted analysis, requiring two large-R jets ($R = 1.0$).

# 2 The Experiment

The LHC accelerates two proton beams in opposite directions around its circumference and collides them at four locations. Particle detectors at these four locations measure the kinematic results of these collisions. These four experiments are ALICE, ATLAS, CMS and LHCb [26–29]. Two of these experiments are specialised: ALICE for heavy ion collisions and LHCb for heavy flavour studies. The other two, ATLAS and CMS, explore a wider area of particle physics.

## 2.1 The ATLAS Detector

The ATLAS experiment at the LHC [27] is a particle detector consisting of inner tracking devices surrounded by a thin superconducting solenoid, electromagnetic and hadronic calorimeters, and a muon spectrometer (shown in Figure 7). Following proton-proton collisions, the inner tracking detector tracks charged particles in pseudorapidity range $|\eta| < 2.5$, a lead/liquid-argon electromagnetic sampling calorimeter covers pseudorapidity $|\eta| < 3.2$ and a steel/scintillator-tile calorimeter covers hadronic energy measurements in the range $|\eta| < 1.7$. The endcap and forward regions have liquid-argon calorimeters for both electromagnetic and hadronic energy measurements up to $|\eta| = 4.9$. The muon spectrometer surrounds the calorimeters and includes *triggering* and tracking chambers.

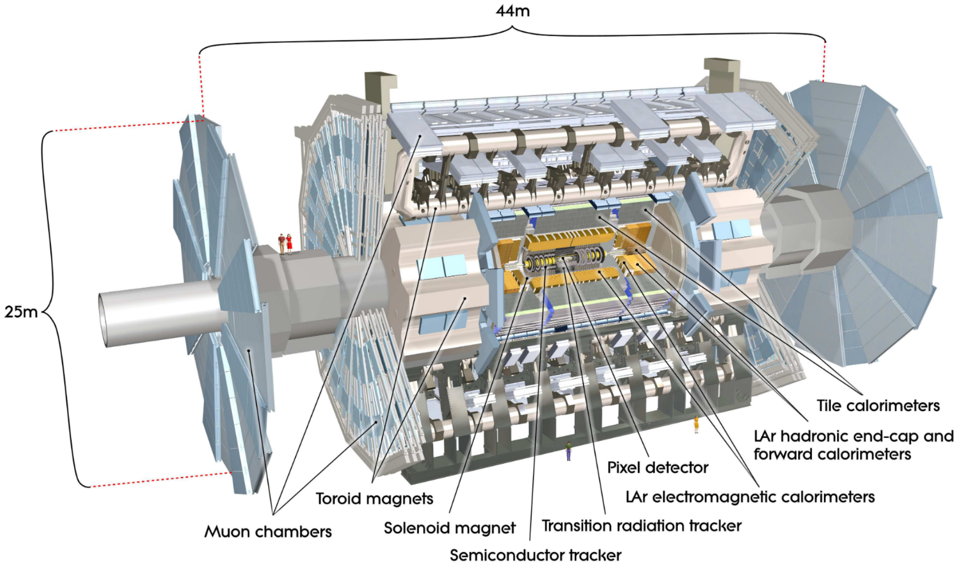

44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

**Figure 7** Schematic of the ATLAS detector and its sub-detectors.

## 2.2 Event Selection

The *triggering* system evaluates the properties of each collision, deciding which event to record. Given that the LHC collides protons every 25 ns, saving data for every bunch crossing is impossible due to information storage and recording rate limitations. Thus, to search for $4b$ signals, $b$-jets must be identified accurately and efficiently. Following the results of trigger optimization studies [19], current $hh \rightarrow 4b$ searches use a dedicated selection of various unprescaled triggers to identify $b$-jets efficiently.

A reconstruction algorithm then estimates the characteristics of each jet using input from the various subsystems of the detector. The most recent improvements in $b$-jet reconstruction come from the introduction of particle flow jets [30], which use tracking information and calorimeter energy deposits to enhance the angular and transverse momentum resolution of jets (see Figure 8), and the new DL1r ATLAS flavour tagging algorithm [31]. In the resolved scheme, low transverse momentum ($p_T$) Higgs bosons produce well-separated $R = 0.4$ anti-$k_t$ $b$-jets. To ensure the trigger is suitably efficient for these events, jets with $p_T > 40$ GeV and $|\eta| < 2.5$ are selected. In search for resolved di-Higgs production, events with at least four $b$-tagged jets are selected, with the four highest $p_t$ jets chosen to correspond to the final state of the $hh \rightarrow 4b$ process.

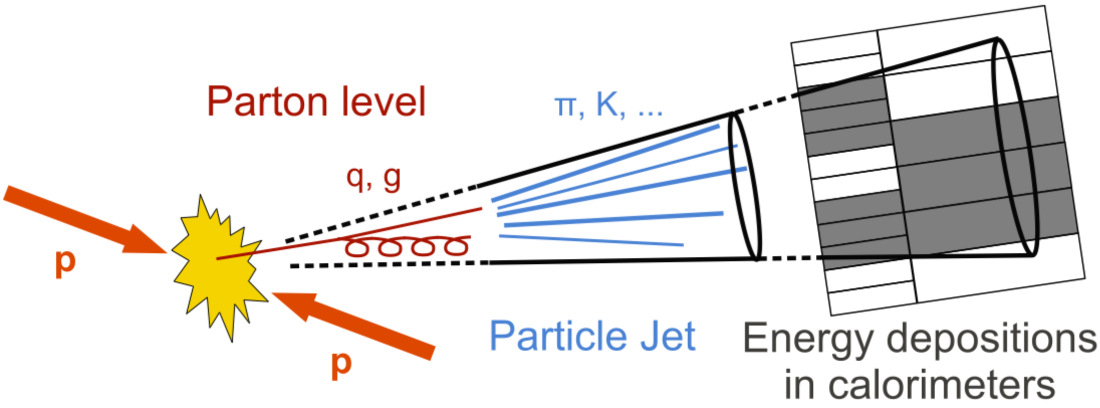


**Figure 8** Schematic of jet reconstruction from calorimeter energy clusters using flavourtagging algorithms.

# 3 Analysis Strategy

The major challenge in identifying $4b$ signals involves modelling two major multi-jet backgrounds, QCD ($\sim 95\%$) and $t\bar{t}$ ($\sim 5\%$), which together are 3-4 orders of magnitude larger than the signal. Since QCD multijet processes are difficult to model with Monte Carlo samples, a data-driven background modelling technique is essential in extracting signals. The main contribution of this project involves devising an alternative strategy for estimating the background in the search for the $4b$ signal via probabilistic deep learning. In particular, *normalizing flows* are explored as a means of estimating the density of $4b$ events, $p_{4b}(\mathbf{x})$, across kinematic variables $\mathbf{x}$.

## 3.1 Jet Pairing

The first step of the analysis requires two *Higgs candidates* $(h_1, h_2)$ to be established for each $4b$ event. Following event selection, there are three ways to pair four jets into two Higgs candidates, as shown in Figure 9. An algorithm is designed to pair the jets most consistently with a di-Higgs topology (the decay of two particles of equal mass). This project considers the vectorised Min($dR$) (MDR-VEC) pairing scheme, which first sorts the two Higgs candidates by the *vector* sum of $p_T$ of their constituent jets (choosing the leading one as $h_1$), then chooses the pairing with the minimum angular separation ($dR$).

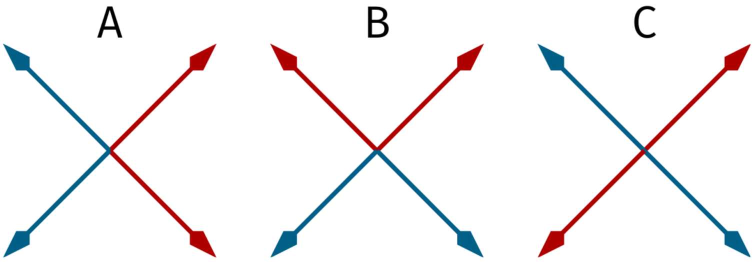


**Figure 9** Three ways of pairing four jets.

The invariant masses of the two Higgs candidates $(m_{h_1}, m_{h_2})$ are constrained to be close to 125 GeV. The *leading* jet mass is associated with $m_{h1}$ while the *sub-leading* jet mass is associated with $m_{h2}$. The $m_{h_1} - m_{h_2}$ plane for the MDR-VEC pairing scheme is shown in Figure 10. It is in this plane that *kinematic regions* are defined and the analysis formulated.

### 3.2 Kinematic Regions

The $4b$ data is separated into two orthogonal datasets[1], the signal region (SR) and control region (CR). The SR is defined by the requirement:

$$X_{hh} = \sqrt{\left(\frac{10(m_{h1} - 120 \text{ GeV})}{m_{h1}}\right)^2 + \sqrt{\left(\frac{10(m_{h2} - 110 \text{ GeV})}{m_{h2}}\right)^2}} < 1.6 \tag{7}$$

The CR consists of those events not in the SR, but pass:

$$\sqrt{(m_{h1} - 1.05 \times 120 \text{ GeV})^2 + (m_{h2} - 1.05 \times 110 \text{ GeV})^2} < 45 \text{ GeV}. \tag{8}$$

The definitions are chosen from a brute-force optimisation procedure conducted in previous iterations of the analysis [19]. $X_{hh}$ is designed to represent the distance of an event from the di-Higgs peak in the $m_{h1}$–$m_{h2}$ plane. The centre of the SR is below $m_h = 125$GeV to account for jet constituents falling outside the jet cone, neutrinos from leptonic $b$ decays, and inactive areas of the detector. The centre of the CR is higher than the SR to avoid the low-mass peak in the background distribution, while also containing a sufficient number of events to estimate the background model. The denominator of $X_{hh}$ represents the resolution of the reconstructed Higgs candidates masses [32].

### 3.3 Background Estimation

The *conditional* probability density of $4b$ events $p_{4b}(\mathbf{x}|\mathbf{m})$, where $\mathbf{x}$ is a vector of kinematic variables and $\mathbf{m}$ is a 2-dimensional vector of the Higgs candidate masses $\mathbf{m} = (m_{h1}, m_{h2})$, is first learnt in the CR, a region of phase-space where the signal contamination is known to be *low*. The signal is then extrapolated to the SR via *inference* (whereby the *learnt* density is conditioned on SR Higgs candidate masses).

[1] In existing analysis [19], the data is split into three orthogonal sets; Control Region (CR), Validation Region (VR) and Signal Region (SR). The VR is a region in-between the CR and SR that is used to quantify systematic uncertainty in the reweighting scheme. A systematic uncertainty has yet to be prescribed for this procedure, so the VR is not used. Instead, the CR in this analysis is a region that would include the VR from previous versions of the analysis.

Comparing the inferred signal with the observed signal allows the background in the SR to be estimated.

Critically, however, the SR is *blinded* in order to make an unbiased optimisation. Therefore, an *alternative* set of SR data is generated for use in inference. This set is provided by a *Gaussian process regression* of CR Higgs candidate mass data, which predicts SR Higgs candidate masses that can be used to model the background before unblinding. Furthermore, an equivalent procedure can be carried out to estimate the background in unblinded $2b$ signals to evaluate performance.

### 3.3.1 Density Estimation

Given a process that independently generates $N$ $D$-dimensional real vectors $\mathbf{x}'$, the *probability density* at $\mathbf{x}' \in \mathbb{R}^D$ measures the extent to which data is generated inside an infinitesimal volume surrounding a vector $\mathbf{x}$. If we let $H_\epsilon(\mathbf{x})$ be a $D$-dimensional hypersphere centered on $\mathbf{x}$, the probability density at $\mathbf{x}$ is given by:

$$p(\mathbf{x}) = \frac{Pr(\mathbf{x}' \in H_\epsilon(\mathbf{x}))}{\mid H_\epsilon(\mathbf{x}) \mid} \quad \text{for } \epsilon \to 0, \tag{9}$$

where $Pr(\mathbf{x}' \in H_\epsilon(\mathbf{x}))$ is a continuous measure of the probability of data being generated inside the hypersphere [33]. A function $p(\cdot)$ taking an arbitrary vector $\mathbf{x}$ and returning the density at $\mathbf{x}$ is called a probability density function. Equipped with the following two properties:

$$p(\mathbf{x}) \geq 0 \quad \text{for all} \quad \mathbf{x} \in \mathbb{R}^D, \tag{10}$$

$$\int_V p(\mathbf{x}) d\mathbf{x} = Pr(\mathbf{x}' \in V), \tag{11}$$

the Radon–Nikodym theorem guarantees that a unique density function exists for any process [34].

This analysis models $hh \rightarrow nb$ decay as a *process* generating a vector of kinematic variables describing each event. The aim is to determine the probability density function of $hh \rightarrow nb$ decay given the kinematics recorded by the ATLAS detector. However, the density function of such physical processes often do not possess simple analytical forms. *Density estimation* encompasses the various algorithms for which the probability density at an arbitrary location $\mathbf{x}$ can be estimated given a set of independently and identically generated datapoints $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ [35].

Density estimation is amongst the most fundamental problems in machine learning: discovering structure from data in an unsupervised manner. A density function effectively contains a complete description of the joint statistical properties of the data. The complexity of phenomena in science and engineering often make probabilistic models the only feasible approach from the computational point of view, even if the underlying processes are inherently deterministic.

The efficacy of density estimation, as with other machine learning tasks, is limited by the dimensionality of our data. *The curse of dimensionality* dramatically increases the difficulty of density estimation in high dimensional space. The issue is illustrated by considering the expected fraction of training datapoints that fall inside the hypersphere $H_\epsilon(\mathbf{x})$. Given that the volume of the hypersphere is:

$$\mid H_\epsilon(\mathbf{x}) \mid = \frac{(\pi^{\frac{1}{2}}\epsilon)^D}{\Gamma(\frac{D}{2}+1)}, \tag{12}$$

where $\Gamma(\cdot)$ is the Gamma function, $\mid H_\epsilon(\mathbf{x}) \mid$ approaches zero for increasing $D$, regardless how large the radius $\epsilon$ is made. Thus, for large $D$, the expected fraction of training datapoints would also approach zero, giving an estimated density of zero. For a reasonable density estimate, a volume large enough to contain at least one datapoint is needed, and so an increasingly large number of datapoints would be required for increasing $D$.

The ATLAS experiment records a large variety of kinematic data for each event. However, with careful assumptions and model design, it is possible to train good density models on high-dimensional data. For example, we can assume the process has fewer degrees of freedom than the measurements we make to describe it. We can select minimally correlated kinematic variables, search for symmetries, constrain the model to lie on a manifold of low intrinsic dimensionality, or perform PCA, t-SNE or any other dimensionality reduction algorithm. Given that the ATLAS experiment recorded $\sim 10^4$ $4b$ events in 2018, a *5-dimensional* vector of minimally correlated kinematic variables is deemed sufficient in constructing a reasonable density estimate. The five variables chosen are: transverse momenta of the two Higgs candidates, $p_T^{h_1}$ and $p_T^{h_2}$, the pseudorapidity of the two Higgs candidates, $\eta_{h_1}$ and $\eta_{h_2}$, and the difference in their azimuthal angle $\Delta\phi_{hh}$.

Methods for density estimation fall into one of two categories: parametric models or non-parametric models. Parametric models assume the density conforms to a specified functional form with a fixed number of adjustable parameters. Examples include simple Gaussian models and mixture models [36, 37], which can be estimated via maximum likelihood or expectation-maximisation respectively. Non-parametric models assume the density only to be smooth and fall into an appropriately restricted, infinite dimensional class of functions. Some non-parametric density estimators are motivated as extensions of the classical histogram, and prove useful for data visualisation, such as violin plots [38]. Others are designed to be more robust to slight changes in the data, such as kernel density estimators [39, 40] and the method of nearest neighbors [41].

However, simple parametric models with few tunable parameters lack flexibility. Mixture models with sufficiently many components can approximate any density arbitrarily well but may require a large number of components to do so. The complexity of non-parametric approaches grows with the number of training datapoints (for which ATLAS has a lot of), and tasks such as inference become more difficult. One promising approach is to design a density estimator that is parametrised by neural networks. Neural networks scale well to large datasets (due to the nature of stochastic gradients), while retaining great flexibility, even allowing us to incorporate domain knowledge in their design. With an appropriately chosen objective function, a neural network can be designed to estimated densities without the drawbacks of traditional methods.

### 3.3.2 Neural Density Estimation

*Neural density estimation* is a parametric method for density estimation that uses neural networks to parameterise a particular density model [33]. A neural density estimator is a neural network with parameters $\phi$ that takes as input a datapoint $\mathbf{x}$ and returns a real number $q_\phi(\mathbf{x})$. Given training data $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, the neural network is typically trained by maximizing the average log likelihood:

$$L(\phi) = \frac{1}{N} \sum_n \log q_\phi(\mathbf{x}_n) \tag{13}$$

The law of large numbers converges $\mathcal{L}(\phi)$ to $\mathbb{E}_{p(\mathbf{x})}(\log q_\phi(\mathbf{x}_n))$, where $\mathbb{E}_{p(\mathbf{x})}$ is the *expectation* with respect to $p(\mathbf{x})$. Hence, for a large enough training set, maximizing $\mathcal{L}(\phi)$ is equivalent to minimizing the Kullback–Leibler divergence between the true density distribution and the estimated density distribution, $D_{KL}(p(\mathbf{x}) \mid\mid q_\phi(\mathbf{x}_n))$:

$$D_{KL}(p(\mathbf{x}) \mid\mid q_\phi(\mathbf{x}_n)) = \mathbb{E}_{p(\mathbf{x})}(\log q_\phi(\mathbf{x}_n)) + \text{const.} \tag{14}$$

Hence a maximum-likelihood density estimator has useful asymptotic properties. In particular, they are asymptotically normally distributed, and asymptotically unbiased and efficient.

Neural density estimators have achieved impressive results in modelling high dimensional data such as natural images [42, 43] and audio data [44]. State-of-the-art neural density estimators have also been used for likelihood-free inference from simulated data [45], variational inference [46], and maximum entropy models [47]. In order to estimate the density of 4$b$ events, a recent class of neural density estimator is used, namely *normalizing flows.*

### 3.3.3 Normalizing Flows

*Normalizing flows* [48] provide a simple mechanism to build expressive density estimators through an iterative procedure. The target distribution is built by applying a chain of $K$ bijective transformations, $f_k$, to a simple prior estimate, $q_0(\cdot)$, with a known and computationally cheap probability density function, such as a Gaussian. Each transformation is parametrised by a neural network, and the last iterate, $q_K(\cdot)$, embodies a more flexible estimate of the target density distribution upon maximizing its average log likelihood. In practice, a random sample $\mathbf{x}_0$ is taken from the prior before being sent through the network, and the output variable has a distribution that matches the one we are interested in. By the chain rule:

$$\mathbf{x}_K = f_K \circ \cdots \circ f_1(\mathbf{x}_0), \quad \mathbf{x}_0 \sim q_0(\mathbf{x}_0). \tag{15}$$

If $f_k$ is a differentiable function with a differentiable inverse, then by a change of variables:

$$\mathbf{x}_K \sim q_K(\mathbf{x}_K) = q_0(\mathbf{x}_0) \prod_{k=1}^{K} \left| \det \frac{\partial f_k}{\partial \mathbf{x}_{k-1}} \right|^{-1}, \tag{16}$$

where $\frac{\partial f_k}{\partial \mathbf{x}_{k-1}}$ is the Jacobian derivative of the transformation. Hence, only transformations with invertibile and efficient Jacobian determinants are considered. Often, $f_k$ is chosen to have an upper or lower triangular Jacobian matrix, given that the determinant of a triangular matrix is simply the product of its diagonals. The log likelihood can then be cheaply evaluated as the sum of two tractable terms:

$$\log q_K(\mathbf{x}_K) = \log\left(q_0\left(f_K^{-1} \circ \cdots \circ f_1^{-1}(\mathbf{x}_0)\right)\right) + \sum_{k=0}^{K} \log \left| \det \frac{\partial f_k}{\partial \mathbf{x}_{k-1}} \right|^{-1} \tag{17}$$

Intuitively, this equation states that the target density is equal to the density at the corresponding point in $\mathbf{x}_0$, plus a term that corrects for the change in volume around an infinitesimally small length around $\mathbf{x}_K$ caused by the transformation. Hence, if we choose $\mathbf{x}_0 \sim q_0(\mathbf{x}_0)$ to be a simple Gaussian, we can build distributions that are complex yet remain easy to sample from. To aid the training procedure, the kinematic variables are preprocessed to $\log p_T^{h_1}$, $\log p_T^{h_2}$, $\eta_{h_1}$, $\eta_{h_2}$ and $\log(\pi - \Delta\phi_{hh})$ such that their distributions look more Gaussian.

A variety of normalizing flows have been proposed, most commonly being parametrised by one of two types of neural networks that guarantee a triangular Jacobian matrix: *affine coupling layers* or *autoregressive layers.* This project considers *neural spline flows* of the *autoregressive* type.

### 3.3.4 Autoregression

A joint density can be decomposed into a product of conditional univariate densities via the chain rule:

$$p(\mathbf{x}) = \prod_i^D p(\mathbf{x}_i|\mathbf{x}_{1:i-1}). \tag{18}$$

An *autoregressive* flow models the joint density by applying a normalizing flow to each univariate density $p(\mathbf{x}_i|\mathbf{x}_{1:i-1})$. Specifically, a normalizing flow is built via a set of functions (neural networks) that can be decomposed as *conditioners* $c^{(i)}$ and *transformers* $t^{(i)}$:

$$\begin{aligned} \mathbf{x}'_i &= f_\theta^{(i)}(\mathbf{x}_{1:i}) && (19) \\ &= t_\theta^{(i)}(\mathbf{x}_i, c_\theta^{(i)}(\mathbf{x}_{1:i})), && (20) \end{aligned}$$

where each transformer $t^{(i)}$ is an invertible function with respect to $\mathbf{x}_i$, and each $c^{(i)}$ is an unrestricted function [49]. Since each $\mathbf{x}'_i$ depends only on $\mathbf{x}_{1:i}$, the Jacobian matrix $J = \frac{\partial \mathbf{x}'}{\partial \mathbf{x}}$ will be lower triangular, which can be cheaply computed by the product of its diagonal values:

$$\det J = \prod_{i=1}^{D} J_{ii} \tag{21}$$

In practice, autoregressive models must impose an order on the variables. Generally, arbitrary orderings can be used.

Autoregressive flows are easily generalized to conditional density estimation. Given a set of example pairs, $\{\mathbf{x}_n, \mathbf{m}_n\}$, the conditional density $p(\mathbf{x} \mid \mathbf{m})$ can be built by simply augmenting the set of input variables with $\mathbf{m}$ and only modelling the conditionals that correspond to $\mathbf{x}$. Simply, this just means the additional inputs are propagated through all hidden layers to all outputs of the $c^{(i)}$ network.

### 3.3.5 Neural Spline Flows

The transformer $t^{(i)}$ must be an invertible and element-wise function of $\mathbf{x}_i$ for the flow as a whole to be invertible. The parameters of the transformation are determined by the conditioner function $c^{(i)}$, a neural network of arbitrary complexity. A family of functions meeting the invertibility requirement are *splines.*

Specifically, we consider $t^{(i)}$ as a *linear rational spline*, a piece-wise function where each of the pieces are linear rational functions of the form $y = \frac{ax+b}{cx+d}$. The spline is specified by the number of intervals (bins) and their locations. Each bin boundary point is called a knot. From a high-level, a spline is a parametrisable curve where knots are joined together. The knots and their derivatives are thus parameters to be determined by $c^{(i)}$.

For knot locations $\left\{\left(x^{(k)},\ y^{(k)}\right)\right\}_{k=0}^{K}$ and bin locations $\left[x^{(k)}, x^{(k+1)}\right]$, where $x^{(k)}$ and $x^{(k+1)}$ are two consecutive points, a continuous linear rational spline is defined by a homographic function of the form:

$$t(x) = \frac{w^{(k)}y^{(k)}(1-\phi) + w^{(k+1)}y^{(k+1)}\phi}{w^{(k)}(1-\phi) + w^{(k+1)}\phi}, \tag{22}$$

where $0 < \phi = \big(x - x^{(k)}\big)/\big(x^{(k+1)} - x^{(k)}\big) < 1$ and $w^{(k)}$ and $w^{(k+1)}$ are two arbitrary weights. [50] proposed an algorithm to construct a *monotone* (and as a result, invertible) formulation of (22). Algorithm 1 from [51] is used to construct $t^{(i)}$ as a *monotone linear rational spline*. The inverse has the same form as its forward form, but with different parameters. Hence, the forward and inverse function evaluations cost the same.

### 3.3.6 Gaussian Process Regression

A *Gaussian process regression* [52] fits a regression function $f(\mathbf{x})$ to the samples of a given training set, which in this case is the full mass plane $\mathbf{m}$ of $nb$ events, excluding the blinded SR. The regression function $f(\mathbf{x})$ is a stochastic process called a *Gaussian process* (GP), which assumes the joint distribution of a finite number of function values $p(f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))$ is itself Gaussian:

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}), \tag{23}$$

where $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))$, $\boldsymbol{\mu} = (m(\mathbf{x}_1), \dots, m(\mathbf{x}_N))$ and $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. $m$ is the mean function and is commonly set to $m(\mathbf{x}) = 0$. $\kappa$ is a positive definite (invertible) kernel function. Thus, Gaussian processes extend multivariate Gaussians to infinite-sized collections of real-valued variables. In particular, Gaussian processes can be considered as distributions over random *functions*, whose shape is defined by $\mathbf{K}$. If points $\mathbf{x}_i$ and $\mathbf{x}_j$ are considered similar by the kernel, $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, are expected to be similar too.

Given training data $\mathbf{X} \in \mathbb{R}^{n\times p}$ and test data $\mathbf{X}^\star \in \mathbb{R}^{m\times p}$, a Gaussian process prior can be converted into a Gaussian process posterior $p(\mathbf{f}^\star|\mathbf{X}^\star, \mathbf{X}, \mathbf{f})$ to make predictions $f^\star$ at new inputs $\mathbf{X}^\star$ . By definition of a Gaussian process, the observations $\mathbf{f}$ and predictions $\mathbf{f}^\star$ are jointly Gaussian:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^\star \end{bmatrix} \sim \mathcal{N}_{n+m}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{K}^\star \\ \mathbf{K}^{\star T} & \mathbf{K}^{\star\star} \end{bmatrix}\right), \tag{24}$$

where $\mathbf{K}^\star = \kappa(\mathbf{X}^\star, \mathbf{X})$ and $\mathbf{K}^{\star\star} = \kappa(\mathbf{X}^\star, \mathbf{X}^\star)$. The predictive distribution is given by:

$$p(\mathbf{f}^\star|\mathbf{X}^\star, \mathbf{X}, \mathbf{f}) = \mathcal{N}(\mathbf{f}^\star|\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star) \tag{25}$$

$$\boldsymbol{\mu}^\star = \mathbf{K}^{\star T}\mathbf{K}^{-1}\mathbf{f} \tag{26}$$

$$\boldsymbol{\Sigma}^\star = \mathbf{K}^{\star\star} - \mathbf{K}^{\star T}\mathbf{K}^{-1}\mathbf{K}^\star \tag{27}$$

The smoothness of the function $f(\mathbf{x})$ determines whether the function overfits or underfits the training data. A function that is not smooth enough may be overly affected by noisy samples, while an overly smooth function may not follow the training samples closely enough. The smoothness is controlled via the kernal function, a common choice of which is the radial basis function (RBF):

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\ell^2}\right), \tag{28}$$

where $\ell$ is the lengthscale hyperparameter determining the smoothness. $\ell$ can be optimised by comparing the Gaussian process prediction of $2b$ signals in the SR with observed $2b$ signals in the SR, to avoid unblinding $4b$ data.

### 3.4 Setting Limits

The corrected di-Higgs invariant mass $m_{hh}^{\text{cor}}$ in the SR is chosen as the discriminating variable to evaluate the upper limit on the production cross section. Following SR inference, histograms of $m_{hh}^{\text{cor}}$ for the observed signal, predicted signal and estimated background are created. The observed signal is compared to the predicted signal and the estimated background to determine how consistent (or significant) the background estimation really is. Either the observed signal is more consistent with the background estimate $B$, or with the predicted signal plus background estimate $S + B$. The agreement is evaluated via a binned maximum-likelihood fit [32].

The maximum likelihood fit is controlled by a signal strength parameter $\mu$, which *scales* the signal histogram to determine how much signal in addition to the estimated background gives the best compatibility to the data, $B + \mu S$. The upper limit is then extracted using the $CL_S$ method [53], with the test statistic:

$$q_\mu = \begin{cases} -2\ln\left(\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}\right) & \hat{\mu} \leq \mu \\ 0 & \hat{\mu} > \mu \end{cases} \tag{29}$$

where $\mu$ is the tested signal strength parameter, $\hat{\mu}$ is the maximum likelihood estimate of the signal strength parameter, $\theta$ is the set of nuisance parameters, $\hat{\theta}$ is the maximum likelihood estimate of the nuisance parameters and $L$ is the profile likelihood. $L(\hat{\mu}, \hat{\theta})$ is the *unconstrained* likelihood and $L(\mu, \hat{\hat{\theta}}(\mu))$ is the *constrained* likelihood, the ratio of which is the most statistically powerful variable to reject one hypothesis in favour of another [37]. $q_\mu$ is used to distinguish between the hypothesis that the data contains signal and background, $S + B$, to that of background only, $B$. Minimising $q_\mu$ coincides with the best compatibility of the data with the $B + \mu S$ hypothesis. $CL_S$, for some test value of $\mu$ is then defined by:

$$CL_S = \frac{CL_{S+B}}{CL_B} = \frac{p(q_\mu \geq q_{\mu,\text{obs}} \mid S + B)}{p(q_\mu \geq q_{\mu,\text{obs}} \mid B)} \tag{30}$$

where the numerator denotes the $p$-value of the $S + B$ hypothesis and the denominator denotes the $1 - p$-value of the $B$ only hypothesis. The signal model is regarded as excluded at a confidence level of $1 - \alpha$ if one finds:

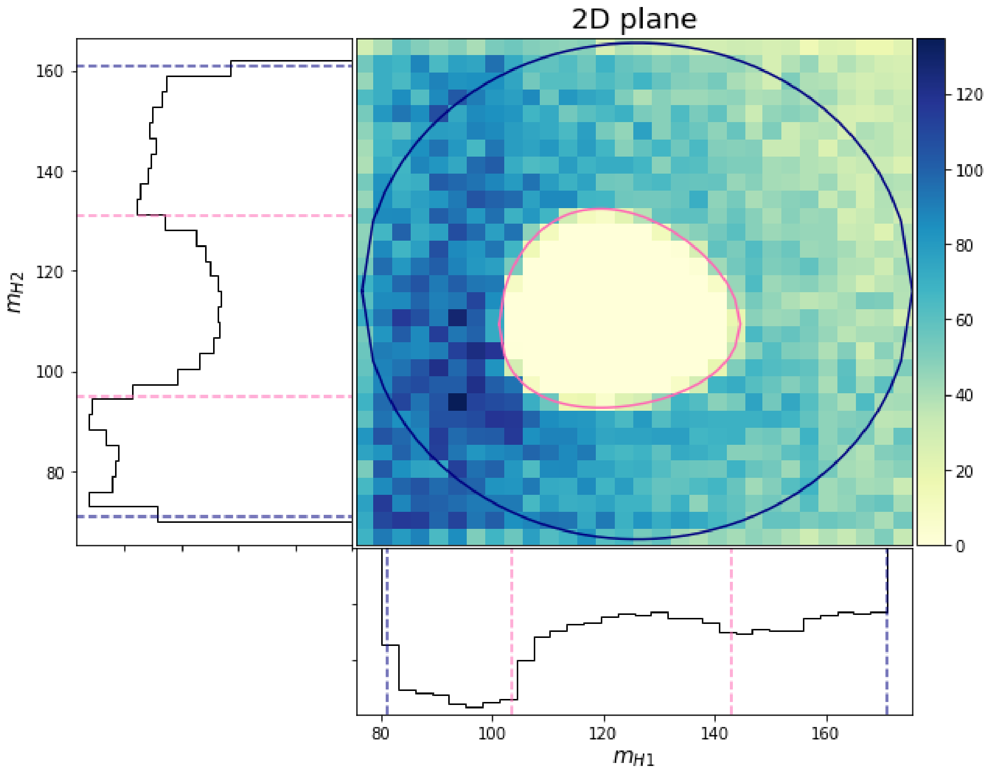
$$CL_S(\mu) < \alpha. \tag{31}$$

Before unblinding, the model is tested on an *Asimov* dataset constructed to model the background estimate in the SR [54]. The Asimov dataset is resembles the background prediction $\mu = 0$ exactly. The *expected* limit is computed with a profile likelihood fit on the Asimov data in place of the SR data, providing a metric for the sensitivity of the analysis, without unblinding the SR. The current analysis has yet to reach the stage to compute observed limits on unblinded data, and so the expected limits are reported.

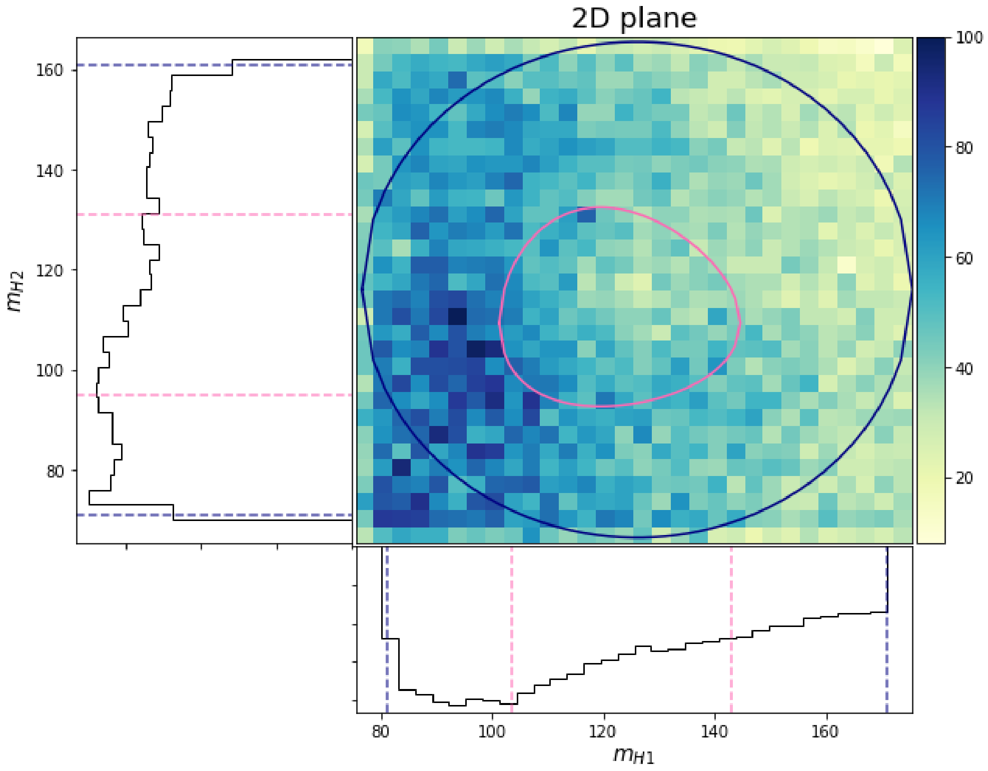
## 4 Analysis Results

The Gaussian process regression was trained for 25 bootstraps on a $(126 - 45)$ GeV $< m_{h_1} <$ $(126 + 45)$ GeV, $(116 - 45)$ GeV $< m_{h_2} <$ $(116 + 45)$ GeV box that enclosed the CR. The training was done on a histogram with $50 \times 50$ bins on the blinded $4b$ mass plane. Bins overlapping with the SR were removed. As shown in Figure 10, the SR prediction demonstrated good closure. Via an equivalent procedure, the regression was validated on a blinded $2b$ mass plane. The $2b$ mass plane was down-sampled to mimic $4b$ statistics. As shown in Figure 11, the regression demonstrated good SR agreement.

The conditional probability density $p_{4b}(\mathbf{x}|\mathbf{m})$ was estimated via three stacked autoregressive neural spline flows trained for 50 epochs and 25 bootstraps on the CR. The transform hyperparameters were optimised by grid search. The autoregression order was randomised. As shown in Figure 12, the probability distribution over the sampled kinematic variables were transformed to give samples over $m_{hh}^{\text{cor, }2}$ and $\cos\theta^{\star}$, where $m_{hh}^{\text{cor, }2} = m_{hh} - m_{h_1} - m_{h_2} + 250$ GeV and $\cos\theta^{\star}$ is the polar angle in the di-Higgs rest frame. The probability density of the SR was evaluated via inference over the Higgs candidate masses predicted by Gaussian process regression. An equivalent procedure was performed on an unblinded, down-sampled $2b$ dataset to validate SR predictions to corresponding distributions in real data, as shown in Figure 13.



**(a)** Blinded $4b$ massplane



**(b)** Gaussian Process massplane

**Figure 10** Diagram depicting the Higgs candidate mass plane (MDR-VEC) before and after Gaussian process regression. Pink and blue contour represents SR and CR respectively.

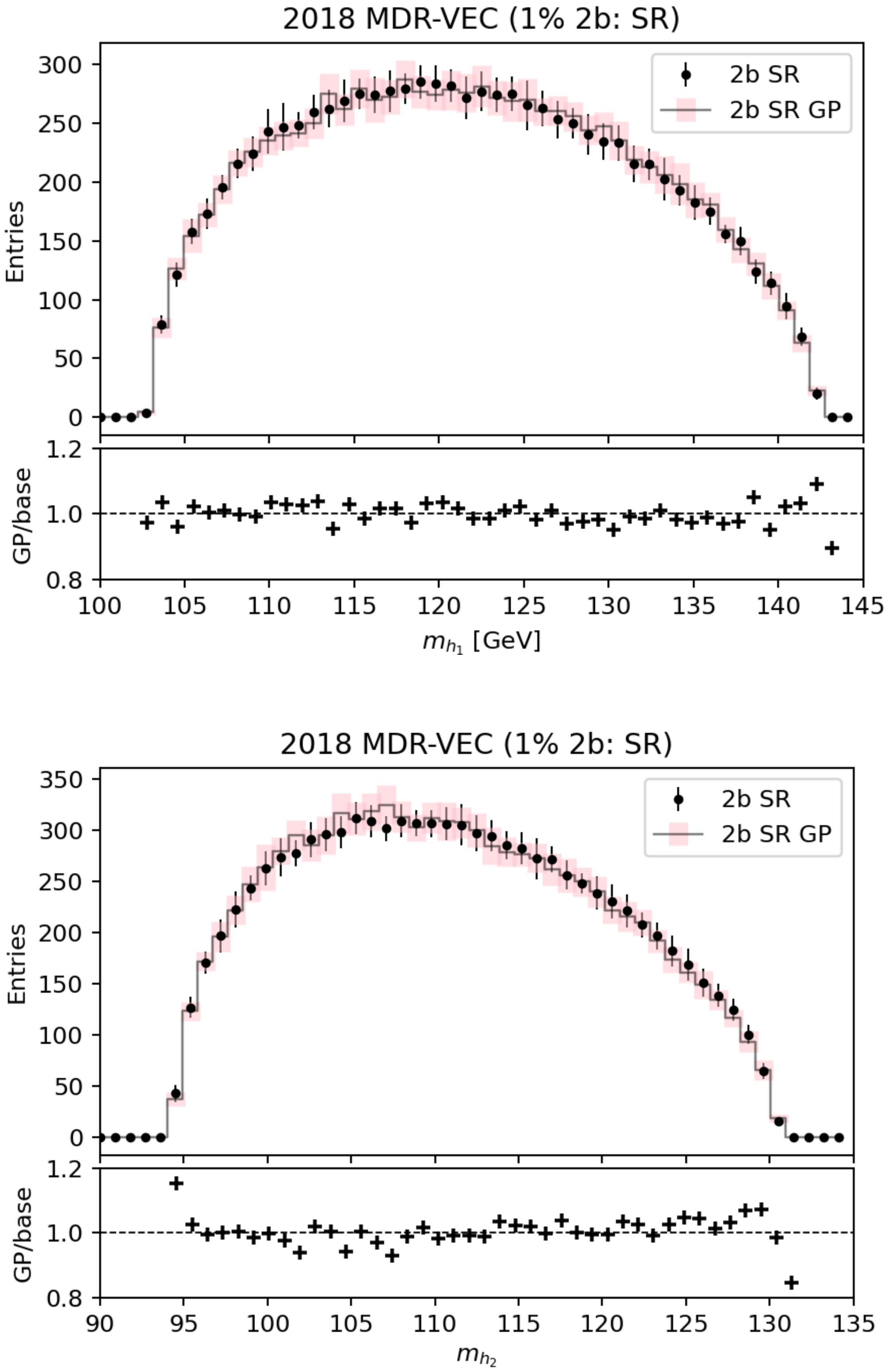


**Figure 11** Validation of Gaussian process regression on down-sampled $2b$ mass plane. Trained for 25 bootstraps.

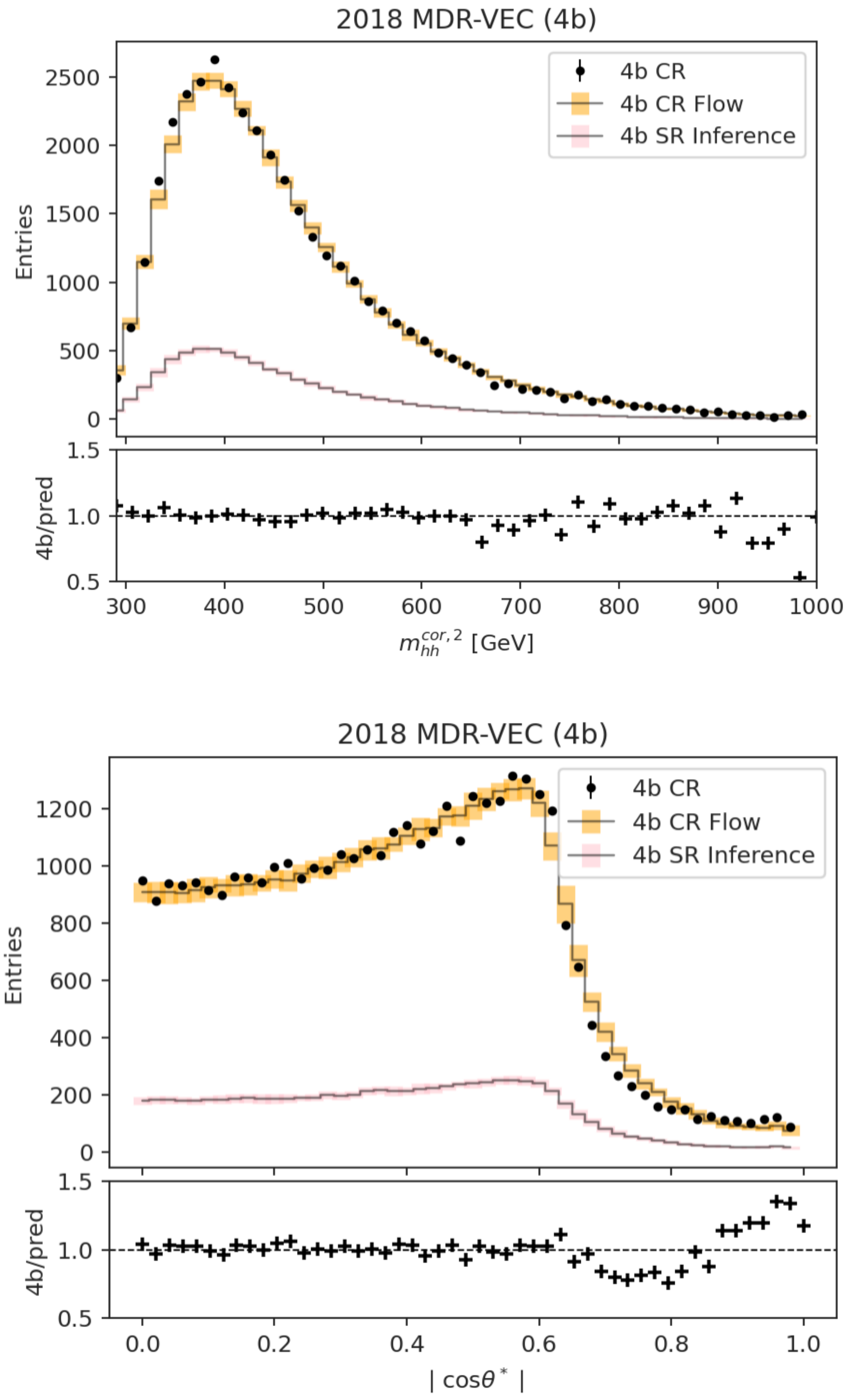


**Figure 12** Density estimation of $4b$ events via conditional neural spline flows. Trained for 25 bootstraps.

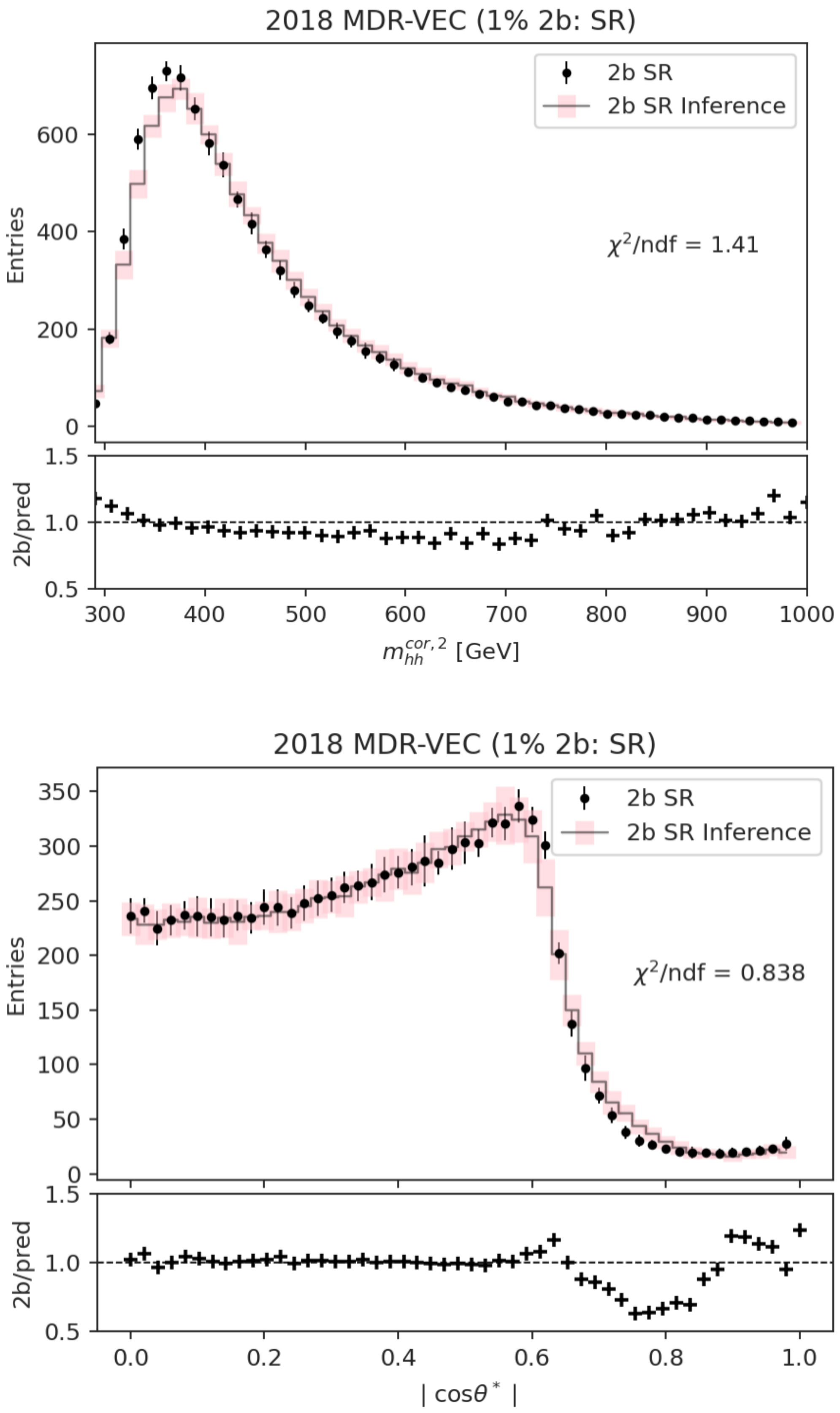


**Figure 13** Validation of flow-based density estimation on down-sampled, unblinded $2b$ data. Trained for 25 bootstraps.

Figure 14 compares the $4b$ SR prediction from flow-based inference to the $4b$ SR prediction from a neural network reweighting of $4b$ CR events. The neural network was trained for 100 bootstraps, reweighting $2b$ CR events to $4b$ CR events [2].

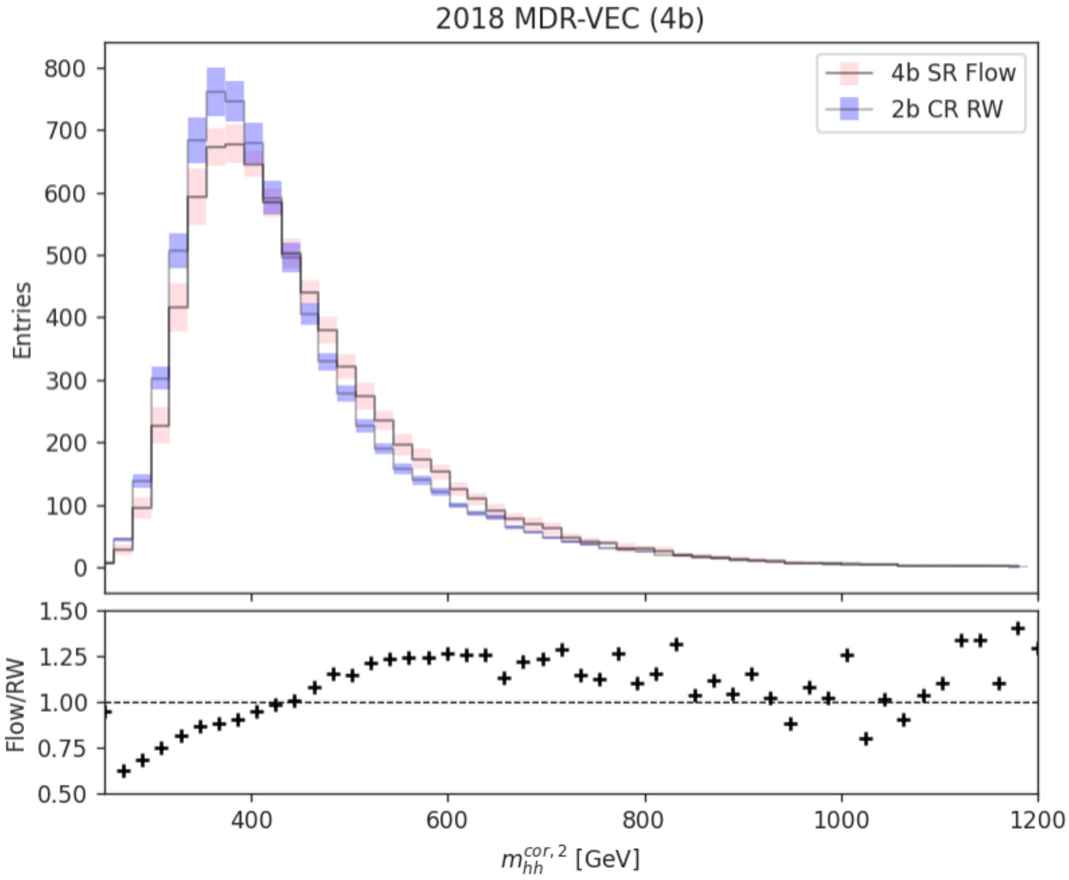


**Figure 14** Comparison of $4b$ predictions from neural network reweighting and flow-based inference.

[2] The definition of the CR differs in the neural network reweighting model due to the inclusion of a VR. Systematic bias is assessed by learning another reweighing function in the VR. The difference between the two background models is taken to be the systematic uncertainty from extrapolating the reweighing function from the CR to the SR. The proposed flow-based model currently has not considered systematic bias.

Table 1 shows the expected 95% $CL_S$ upper limits on the production cross section $\sigma(pp \to hh \to b\bar{b}b\bar{b})$, found by carrying out the $CL_S$ calculation at many $\mu$ test values to determine that at which $CL_S = 0.05$. The one and two sigma uncertainties were determined by additionally calculating $CL_S$ at $\pm 1\sigma$ and $\pm 2\sigma$ at each test $\mu$, and determining where each of these is equal to 0.05. Due to the use of signals normalized to 1fb, $\mu$ is also the signal cross-section in fb.

| -2$\sigma$ | -1$\sigma$ | $\mu$ | +1$\sigma$ | +2$\sigma$ |
|---|---|---|---|---|
| 5.50 | 7.40 | 10.3 | 14.4 | 19.4 |

**Table 1:** Expected 95% CL exclusion limits for SM non-resonant di-Higgs production, in units of the SM prediction for $\sigma(pp \to hh \to b\bar{b}b\bar{b})$. Observed limits yet to be shown. SR remains blinded.

The equivalent limit for the neural network reweighting procedure was $\mu = 10.8$ fb. Thus, the proposed model offered a slight improvement in sensitivity.

## 5 Conclusion & Future Studies

A search for non-resonant production of pairs of Standard Model Higgs bosons has been detailed for the $b\bar{b}b\bar{b}$ channel using $s = \sqrt{13}$ TeV pp collision data collected by the ATLAS detector in 2018. An end-to-end interpolation using neural spline flows and Gaussian process regression is proposed as an alternative strategy to estimate background in $b\bar{b}b\bar{b}$ events. Expected upper limits on the production cross section to the $b\bar{b}b\bar{b}$ final state are reported, demonstrating a slight improvement in sensitivity over the baseline background estimation procedure using neural network reweighting.

While demonstrating a slight improvement on the statistics only limit, the proposed interpolation scheme for background estimation is not expected to replace the current neural network reweighting model. Its performance is highly dependent on the jet pairing algorithm used, producing poor results for Higgs candidate mass planes that have large peaks in the signal region. The flow-based density estimation also struggles to model the invariant di-Higgs mass in the bulk region $320 - 400$ GeV. Systematic uncertainties have yet to be assigned, and the bootstrap error seems to suggest the training has larger instability in comparison to the reweighting model. Instead, the proposed background estimation scheme serves to demonstrate the application of state-of-the-art probabilistic deep learning in high energy physics while also providing a useful cross check for the current background estimation procedure.

With a rapidly developing landscape in deep learning, state-of-the-art networks are an exciting tool to aid the search for new physics. In addition to assigning systematic uncertainty and propagating Gaussian process regression errors forward into the full background estimate, various other improvements could be implemented to the proposed background estimation scheme. The analysis only used five kinematic variables to model the background, in comparison to the eleven used in the reweighting model. Introducing more kinematic variables would aid the sensitivity of the model. A natural extension to normalizing flows is to use them to estimate the probability density of latent variables on a lower dimensional manifold. By introducing latent variables, the global features $nb$ events can be captured, structured and clustered, even for extremely large input dimensionality. A flow-based variational autoencoder is an example of such a model designed to learn structured lower dimensional distributions which contain all the information about the input variables.

At present, investigating di-Higgs production remains a promising avenue in discovering physics Beyond the Standard Model. As the LHC continues to gather more data, searches will gain more sensitivity and will be put to the test once again.

# References

1. Aad, G. *et al.* Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716,** 1–29. arXiv: `1207.7214 [hep-ex]` (2012).

2. Feynman, R. P. Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction. *Phys. Rev.* **80,** 440–457. `https://link.aps.org/doi/10.1103/PhysRev.80.440` (3 Nov. 1950).

3. Weinberg, S. A Model of Leptons. *Phys. Rev. Lett.* **19,** 1264–1266. `https://link.aps.org/doi/10.1103/PhysRevLett.19.1264` (21 Nov. 1967).

4. Glashow, S. L. Partial Symmetries of Weak Interactions. *Nucl. Phys.* **22,** 579–588 (1961).

5. Ferrari, A. *Searches for Higgs boson pair production with ATLAS* in *13th Conference on the Intersections of Particle and Nuclear Physics* (Sept. 2018). arXiv: `1809.08870 [hep-ex]`.

6. Englert, F. & Brout, R. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.* **13,** 321–323. `https://link.aps.org/doi/10.1103/PhysRevLett.13.321` (9 Aug. 1964).

7. Higgs, P. W. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.* **13,** 508–509. `https://link.aps.org/doi/10.1103/PhysRevLett.13.508` (16 Oct. 1964).

8. Bergerhoff, B. & Wetterich, C. Electroweak Phase Transition in the Early Universe? *NATO Sci. Ser. C* **511** (eds Sánchez, N. & Zichichi, A.) 211–240 (1998).

9. Carena, M., Liu, Z. & Riembau, M. Probing the electroweak phase transition via enhanced di-Higgs boson production. *Phys. Rev. D* **97,** 095032. `https://link.aps.org/doi/10.1103/PhysRevD.97.095032` (9 May 2018).

10. Cline, J. M. *Baryogenesis* in *Les Houches Summer School - Session 86: Particle Physics and Cosmology: The Fabric of Spacetime* (Sept. 2006). arXiv: `hep-ph/0609145`.

11. Grazzini, M. *et al.* Higgs boson pair production at NNLO with top quark mass effects. *JHEP* **05,** 059. arXiv: `1803.02463 [hep-ph]` (2018).

12. *Constraints on the Higgs boson self-coupling from the combination of single-Higgs and double-Higgs production analyses performed with the ATLAS experiment* tech. rep. ATLAS-CONF-2019-049 (CERN, Geneva, Oct. 2019). `https://cds.cern.ch/record/2693958`.

13. Branco, G. *et al.* Theory and phenomenology of two-Higgs-doublet models. *Physics Reports* **516,** 1–102. ISSN: 0370-1573. `http://dx.doi.org/10.1016/j.physrep.2012.02.002` (July 2012).

14. Hespel, B., Lopez-Val, D. & Vryonidou, E. Higgs pair production via gluon fusion in the Two-Higgs-Doublet Model. *JHEP* **09,** 124. arXiv: `1407.0281 [hep-ph]` (2014).

15. Randall, L. & Sundrum, R. Large Mass Hierarchy from a Small Extra Dimension. *Phys. Rev. Lett.* **83,** 3370–3373. `https://link.aps.org/doi/10.1103/PhysRevLett.83.3370` (17 Oct. 1999).

16. Dugan, M. J., Georgi, H. & Kaplan, D. B. Anatomy of a Composite Higgs Model. *Nucl. Phys. B* **254,** 299–326 (1985).

17. Bellazzini, B., Csaki, C., Hubisz, J., Serra, J. & Terning, J. Composite Higgs Sketch. *JHEP* **11,** 003. arXiv: `1205.4032 [hep-ph]` (2012).

18. Dorsner, I., Fajfer, S., Kamenik, J. F. & Kosnik, N. Light colored scalars from grand unification and the forward-backward asymmetry in t t-bar production. *Phys. Rev. D* **81,** 055009. arXiv: `0912.0972 [hep-ph]` (2010).

19. Aaboud, M. *et al.* Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics* **2019.** ISSN: 1029-8479. `http://dx.doi.org/10.1007/JHEP01(2019)030` (Jan. 2019).

20. Aad, G. *et al.* Search for Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state from pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *The European Physical Journal C* **75.** ISSN: 1434-6052. `http://dx.doi.org/10.1140/epjc/s10052-015-3628-x` (Sept. 2015).

21. Aaboud, M. *et al.* Search for Higgs boson pair production in the $b\bar{b}WW^*$ decay mode at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics* **2019.** ISSN: 1029-8479. `http://dx.doi.org/10.1007/JHEP04(2019)092` (Apr. 2019).

22. Aaboud, M. *et al.* Search for Resonant and Nonresonant Higgs Boson Pair Production in the $b\bar{b}\tau^+\tau^-$ Decay Channel in *pp* Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector. *Physical Review Letters* **121.** ISSN: 1079-7114. `http://dx.doi.org/10.1103/PhysRevLett.121.191801` (Nov. 2018).

23. Aaboud, M. *et al.* Search for Higgs boson pair production in the $WWW^*WW^*$ decay channel using ATLAS data recorded at $\sqrt{s} = 13$. *Journal of High Energy Physics* **2019.** ISSN: 1029-8479. `http://dx.doi.org/10.1007/JHEP05(2019)124` (May 2019).

24. Aaboud, M. *et al.* Search for Higgs boson pair production in the $\gamma\gamma b\bar{b}$ final state with 13 TeV pp collision data collected by the ATLAS experiment. *Journal of High Energy Physics* **2018.** ISSN: 1029-8479. `http://dx.doi.org/10.1007/JHEP11(2018)040` (Nov. 2018).

25. Aaboud, M. *et al.* Search for Higgs boson pair production in the $\gamma\gamma WW^*$ channel using *pp* collision data recorded at $\sqrt{s} = 13$ TeV with the ATLAS detector. *The European Physical Journal C* **78.** ISSN: 1434-6052. `http://dx.doi.org/10.1140/epjc/s10052-018-6457-x` (Dec. 2018).

26. Aamodt, K. *et al.* The ALICE experiment at the CERN LHC. *JINST* **3,** S08002 (2008).

27. Aad, G. *et al.* The ATLAS Experiment at the CERN Large Hadron Collider. *JINST* **3,** S08003 (2008).

28. Chatrchyan, S. *et al.* The CMS Experiment at the CERN LHC. *JINST* **3,** S08004 (2008).

29. Alves Jr., A. A. *et al.* The LHCb Detector at the LHC. *JINST* **3,** S08005 (2008).

30. Aaboud, M. *et al.* Jet reconstruction and performance using particle flow with the ATLAS Detector. *The European Physical Journal C* **77.** ISSN: 1434-6052. `http://dx.doi.org/10.1140/epjc/s10052-017-5031-2` (July 2017).

31. Aad, G. *et al.* ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV. *Eur. Phys. J. C* **79,** 970. arXiv: `1907.05120 [hep-ex]` (2019).

32. Paredes Saenz, S. R. *Search for New Physics through Higgs-Boson-Pair Production at the LHC and Beyond* Presented 03 07 2020 (June 2020). https://cds.cern.ch/record/2752606.

33. Papamakarios, G. *Neural Density Estimation and Likelihood-free Inference* 2019. arXiv: 1910.13233 [stat.ML].

34. Kramer, W. Probability & Measure : Patrick Billingsley (1995): (3rd ed.). New York : Wiley, ISBN 0-471-0071-02, pp 593, [pound sign] 49.95. *Computational Statistics & Data Analysis* **20,** 702–703. https://ideas.repec.org/a/eee/csdana/v20y1995i6p703-702.html (Dec. 1995).

35. Silverman, B. W. *Density estimation for statistics and data analysis* https://cds.cern.ch/record/1070306 (Chapman and Hall, London, 1986).

36. Bilmes, J. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models* tech. rep. (1998).

37. Lindsay, B. G. Mixture Models: Theory, Geometry and Applications. *NSF-CBMS Regional Conference Series in Probability and Statistics* **5,** i–163. ISSN: 19355920, 23290978. http://www.jstor.org/stable/4153184 (1995).

38. Hintze, J. L. & Nelson, R. D. Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician* **52,** 181–184 (1998).

39. Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics* **27,** 832–837. https://doi.org/10.1214/aoms/1177728190 (1956).

40. Parzen, E. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33,** 1065–1076. https://doi.org/10.1214/aoms/1177704472 (1962).

41. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **46,** 175–185. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00031305.1992.10475879. https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879 (1992).

42. Dinh, L., Sohl-Dickstein, J. & Bengio, S. *Density estimation using Real NVP* in (2017). https://arxiv.org/abs/1605.08803.

43. Rippel, O. & Adams, R. P. *High-Dimensional Probability Estimation with Deep Density Models* 2013. arXiv: 1302.5125 [stat.ML].

44. Van den Oord, A. *et al. WaveNet: A Generative Model for Raw Audio* in *Arxiv* (2016). https://arxiv.org/abs/1609.03499.

45. Papamakarios, G. & Murray, I. *Fast* $\epsilon - freeInferenceofSimulationModelswithBayesianConditionalDensityEstimation$ in *Advances in Neural Information Processing Systems* (eds Lee, D., Sugiyama, M., Luxburg, U., Guyon, I. & Garnett, R.) **29** (Curran Associates, Inc., 2016). https://proceedings.neurips.cc/paper/2016/file/6aca97005c68f1206823815f66102863-Paper.pdf.

46. Rezende, D. & Mohamed, S. *Variational Inference with Normalizing Flows* in *Proceedings of the 32nd International Conference on Machine Learning* (eds Bach, F. & Blei, D.) **37** (PMLR, Lille, France, July 2015), 1530–1538. http://proceedings.mlr.press/v37/rezende15.html.

47. Loaiza-Ganem, G., Gao, Y. & Cunningham, J. P. *Maximum Entropy Flow Networks* 2017. arXiv: 1701.03504 [stat.ME].

48. Kobyzev, I., Prince, S. & Brubaker, M. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1–1. ISSN: 1939-3539. http://dx.doi.org/10.1109/TPAMI.2020.2992934 (2020).

49. Papamakarios, G., Pavlakou, T. & Murray, I. *Masked Autoregressive Flow for Density Estimation* 2018. arXiv: 1705.07057 [stat.ML].

50. Fuhr, R. D. & Kallay, M. Monotone linear rational spline interpolation. *Computer Aided Geometric Design* **9,** 313–319. ISSN: 0167-8396. https://www.sciencedirect.com/science/article/pii/016783969290038Q (1992).

51. Dolatabadi, H. M., Erfani, S. & Leckie, C. *Invertible Generative Modeling using Linear Rational Splines* 2020. arXiv: 2001.05168 [stat.ML].

52. Williams, C. K. & Rasmussen, C. *Gaussian Processes for Regression* in *NIPS* (1995).

53. Read, A. L. Presentation of search results: The CL(s) technique. *J. Phys. G* **28** (eds Whalley, M. R. & Lyons, L.) 2693–2704 (2002).

54. Cowan, G., Cranmer, K., Gross, E. & Vitells, O. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C* **71.** [Erratum: Eur.Phys.J.C 73, 2501 (2013)], 1554. arXiv: 1007.1727 [physics.data-an] (2011).