

DISSERTATION

---

**End to end optimization in a search for  
boosted Higgs boson pair production in  
the  $bbbb$  final state via  
vector-boson-fusion (VBF) production  
using the run 2 dataset with the ATLAS  
detector**

---

For the attainment of the academic degree doctor rerum naturalium

(Dr. rer. nat.) in the subject: Physics

**M.Sc. Frederic Renner**

Berlin, 30.06.2023

Faculty of Mathematics and Natural Sciences of the Humboldt  
University of Berlin

1st Supervisor: Dr. Clara Elisabeth Leitgeb

2nd Supervisor: Prof. Dr. Cigdem Issever

---

(Only after the disputation for publication in the university library according to § 15 of the doctoral regulations enter the names and the date):

Reviewers:

1st:

2nd:

3rd:

Date of the oral examination:



## Abstract

I am an abstract.



---

# Contents

<b>1</b>	<b>Theory</b>	<b>1</b>
1.1	Feynman rules from field theory . . . . .	1
1.2	Probability of a process . . . . .	2
1.3	The Standard Model . . . . .	3
1.4	Statistics . . . . .	4
1.4.1	Building the likelihood . . . . .	4
1.4.2	From test statistic to p-value . . . . .	5
1.4.3	The $CL_s$ value . . . . .	7
1.4.4	The HistFactory model . . . . .	8
	<b>Bibliography</b>	<b>11</b>

---

## List of Figures





---

# List of abbreviations

**ATLAS**    The Atlas Experiment  
**pdf**        probability density function



# Chapter 1

## Theory

Where to start? Sometimes it seems like that particle physics is bringing it all together, as it tries to give a comprehensive picture of the world by describing the structure of matter from quantum mechanics to cosmology. So I would say to start shallow (since we are experimentalists) at the very beginning, and then dive a bit deeper into Standard Model to have a plausible thread how the need and the development of the Higgs mechanism came about. The following is mainly based on [1, 2] and intended to make the calculation of cross sections plausible.

### 1.1 Feynman rules from field theory

The fact that elementary particles can seemingly be born out of nothing and die again led to the development of their currently most successful description through quantum field theories. Heuristically it can be understood by the uncertainty principle, which states that energy can vary greatly on short time scales, and by special relativity, which allows the property energy to be converted into the property mass. This marriage between quantum mechanics and special relativity is what drove the development of quantum field theory.

can be deduced through perturbation theory, just on term, The conventional strategy is perturbation theory with the free fields as starting point, treating the interaction as a small perturbation

To make a field one assigns a quantity to some region in spacetime, e.g.  $\phi(\mathbf{x}, t)$ . A Lagrangian  $L(\phi(\mathbf{x}, t))$  then governs the dynamics, like excitations or interactions of this field, which can e.g. represent the birth and death of particles or interactions by the exchange of a particle between them. One formulation of quantum field theory is by use of the path integral formulation. It then basically boils down to integrals of the form  $\int D\phi e^{i \int d^4x L(\phi(\mathbf{x}, t))}$ . Where  $\int D\phi$  is the integral over all possible paths/ways a particle could take. Through back and forth expansions of the  $e$  functions the integral can be solved and the result is a probability - the amplitude  $\mathcal{M}$  of e.g. an interaction between two particles, like scattering, usually depicted in the form of Feynman Diagrams. As this follows a pattern the formalism can be contracted into the infamous Feynman rules (for details see [2]).

## 1.2 Probability of a process

Probes of elementary particle interactions are accessible via bound states, decays and scattering. The first can be studied within classical quantum mechanics whereas the latter uses the preceding. Since this work deals with a collider experiment I think its at least useful to see how one can calculate in principle a cross section  $\sigma$ . It is a measure of how possible an interaction is when shooting something at each other. Calculating reaction rates in quantum mechanics is done by Fermi's golden rule. Here the relativistic version for a scattering process like  $1 + 2 \rightarrow 3 + 4 + \dots + n$  is given [2]

$$\sigma = \frac{S\hbar^2}{4\sqrt{(p_1 \cdot p_2)^2 - (m_1 m_2 c^2)^2}} \int |\mathcal{M}|^2 (2\pi)^4 \delta^4(p_1 + p_2 - p_3 \dots - p_n) \times \prod_{j=3}^n 2\pi \delta(p_j^2 - m_j^2 c^2) \Theta(p_j^0) \frac{d^4 p_j}{(2\pi)^4}. \quad (1.2.1)$$

$S$  is a statistical factor accounting for identical particles (e.g.  $a \rightarrow b + b + c + c + c$ , then  $S = (1/2!)(1/3!)$ ),  $p_i$  are four momenta of particle  $i$  over which one integrates,  $\mathcal{M}(p_1, \dots, p_n)$  is the amplitude of the process calculable with the Feynman rules, the  $\delta^4$  ensures energy and momentum conservation, the last  $\delta$  ensures that particles are on their mass shell ( $E_j^2/c^2 - \mathbf{p}_j^2 = m_j^2 c^2$ ) and the Heaveside  $\Theta$  makes sure that

outgoing energies are positive  $p_j^0 = E_j/c > 0$ . With this, a particle physicist can calculate the probability of any process at a collider experiment.

## 1.3 The Standard Model

### 1.1

dirac, require local gauge invariance -> qed Lagrangian

gauge field blah only about Lagrangians, no field solutions needed

## 1.4 Statistics

Every scientific investigation starts with a hypothesis that is to be tested empirically. The main objective is to evaluate if the proposed hypothesis agrees or disagrees with observed data, to either accept or reject it against the null-hypothesis. The metric at hand to do so is the p-value that arises within hypothesis testing.

In the field of high-energy physics, a framework based on likelihood statistics has been developed specifically for this task. This section begins to lay out the mathematical fundamentals of the approach and then goes to the hands on implementation of its use. The following is based on [3–5].

### 1.4.1 Building the likelihood

The statistical model must take into account how compatible the observed collision events are with the theoretical predictions. This can be described by a likelihood  $L(\mathbf{x}|\phi)$  that is just a probability for an observation  $\mathbf{x}$  under a given set of parameters  $\phi$ . Since we are dealing with a counting experiment the tool at hand will be bins of histograms  $\mathbf{h} = (h_1, \dots, h_N)$ .

It is useful to subdivide a measurement  $\mathbf{x} = (\mathbf{n}, \mathbf{a})$  further into a quantity of interest  $\mathbf{n}$ , (e.g. the invariant mass of a particle) and *auxiliary* measurement histograms  $\mathbf{a}$  that help to constrain the model. Additionally in the context of hypothesis testing it is useful to split the set of parameters  $\phi = (\psi, \Theta)$  into so called parameters of interest  $\psi$  and nuisance parameters  $\Theta$ . For this section the parameter of interest will be only one parameter, the signal strength  $\mu$ .

The bin heights (counts) can then be expressed in terms of the amount of signal  $s_i(\Theta)$  and background  $b_i(\Theta)$  in them depending in some way on the nuisance parameters. The prediction (expectation value) of the  $n_i$  can then be expressed as

$$\langle n_i(\mu, \Theta) \rangle = \mu s_i(\Theta) + b_i(\Theta), \quad (1.4.1)$$

and similar for the  $a_i$

$$\langle a_i(\Theta) \rangle = u_i(\Theta). \quad (1.4.2)$$

As we are expecting the bin counts to occur with a constant mean rate and independent of time compared to the last event, each bin follows a Poisson distribution

$$\frac{r^k e^{-r}}{k!}. \quad (1.4.3)$$

$r$  is the expected rate of occurrences, which translates as our prediction, whereas  $k$  are the actual measured occurrences. Accounting for all the bins by multiplying them together yields

$$L(\mu, \Theta) = \prod_{j=1}^N \frac{(\mu s_j(\Theta) + b_j(\Theta))^{n_j}}{n_j!} e^{-(\mu s_j(\Theta) + b_j(\Theta))} \prod_{k=1}^M \frac{u_k(\Theta)^{a_k}}{a_k!} e^{-u_k(\Theta)}. \quad (1.4.4)$$

To test for a hypothesized value of  $\mu$ , the best choice according to the Neyman-Pearson lemma, is the profile likelihood ratio that reduces the dependence to the parameter(s) of interest

$$\lambda(\mu) = \frac{L(\mu, \hat{\Theta})}{L(\hat{\mu}, \hat{\Theta})} \quad (1.4.5)$$

The denominator is the unconditional maximum likelihood estimation so that  $\hat{\mu}$  and  $\hat{\Theta}$  both are free to vary to maximize  $L$ , whereas the numerator is the found maximum likelihood conditioned on some chosen  $\mu$  and the set nuisance parameters  $\hat{\Theta}$  that maximize the likelihood for that particular  $\mu$ . This definition gives  $0 \leq \lambda \leq 1$ . For a  $\lambda \approx 1$  the hypothesized value of  $\mu$  shows good agreement to the Poissonian model.

### 1.4.2 From test statistic to p-value

Transforming the profile likelihood into a test statistic  $t_\mu$  is practical to calculate p-values

$$t_\mu = -2 \log \lambda(\mu). \quad (1.4.6)$$

This translates as  $t_\mu \rightarrow 0$  as good agreement,  $t_\mu \rightarrow \infty$  as bad agreement to the model. A right-tail p-value can then be calculated from the probability density

function of  $t_\mu$ :  $\text{pdf}(t_\mu) = f(t_\mu | \mu)$

$$p_\mu = \int_{t_{\mu,obs}}^{\infty} f(t_\mu | \mu) dt_\mu \quad (1.4.7)$$

$t_{\mu,obs}$  is the test statistic  $t_\mu$  evaluated at the observed data. This is like plugging into the Poisson distributions the same values for  $r$  as for  $k$  in eq. 1.4.3. Just like a probability density function for a standard normal distribution, intuitively the pdf is just, how probable a particular value of the test statistic  $t_\mu$  is under a fixed value of the signal strength (how often it occurs compared to all other values  $t_\mu$  can have).

This particular form is handy because there exist approximations for  $f(t_\mu | \mu)$  [3]. Wald [6] proved that in the large sample limit the test statistic follows a normalized sum of squared distances between the tested parameter of interest  $\mu_i$  and its maximum likelihood estimate  $\hat{\mu}_i$ . The result was extended by Wilk [7] for any number of parameters of interest so the test statistic becomes

$$t_\mu = \sum_i \frac{(\mu_i - \hat{\mu}_i^2)}{\sigma_i^2} + \mathcal{O}(1/\sqrt{N}). \quad (1.4.8)$$

The  $\hat{\mu}_i$  are in the large sample limit normally distributed with mean  $\mu'$  (true values) and standard deviation  $\sigma_i$ . This is the definition of a non-central chi-squared distribution with degrees of freedom  $i$  equal to the parameters of interest (see section 3.1 in [3]). For one parameter of interest the distribution reads

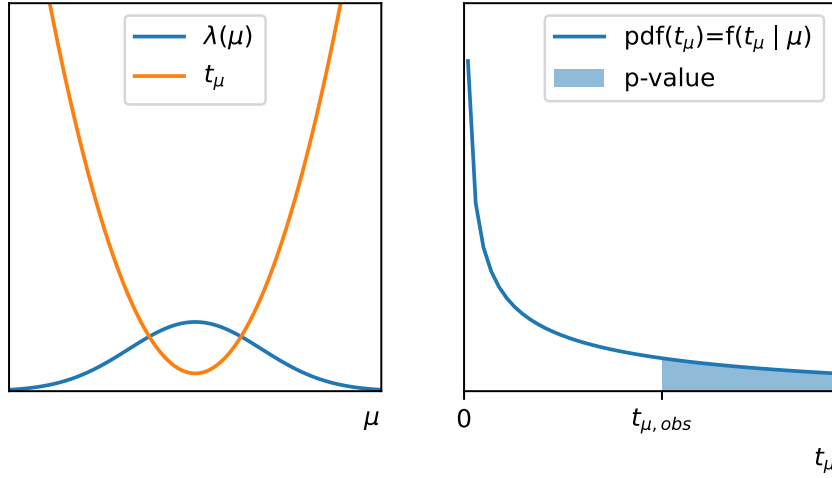
$$f(t_\mu | \mu) = \frac{1}{2\sqrt{t_\mu}} \frac{1}{\sqrt{2\pi}} \left[ \exp\left(-\frac{1}{2}(\sqrt{t_\mu} + \sqrt{\Lambda_\mu})\right) + \exp\left(-\frac{1}{2}(\sqrt{t_\mu} - \sqrt{\Lambda_\mu})\right) \right], \quad (1.4.9)$$

with non-centrality parameter

$$\Lambda_\mu = \frac{(\mu - \mu')^2}{\sigma^2}. \quad (1.4.10)$$

Figure 1.1 illustrates the different steps. Being able to calculate p-values allows now to state how likely it is that the proposed hypothesis is reflected by the observed data. Put differently, if the experiment would be repeated the p-value represents





**Figure 1.1:** A sketch to follow the steps to calculate p-values. (**left**) The profile likelihood (■) has essentially some hill-like form with a maximum at  $\lambda(\hat{\mu}, \hat{\Theta})$ ,  $t_\mu$  (■) is  $-2\ln(\lambda)$ . (**right**) For one parameter of interest in the large sample limit  $f(t_\mu | \mu)$  follows a non-central chi-squared distribution with one degree of freedom, equation 1.4.9. The blue shaded area under the pdf is a right hand sided p-value.

the probability of that the outcome favors the alternative hypothesis over the null hypothesis.

In the scientific community a widely accepted threshold for this is a p-value of 0.05. Though particle physicists only claim discovery of a new phenomenon for  $p < 2.87 \times 10^{-7}$  (5 standard deviations of the standard normal distribution). One caveat here is that this particular form of  $t_\mu$  assumes  $\mu$  can also be negative, which can be non-physical depending on the impact of a new process. Test statistics and their pdf approximations considering the different cases are covered in [3].

### 1.4.3 The $\text{CL}_s$ value

Particle physicists are usually interested in two things when making statistical tests for discovery of new phenomena: how well is the modeling of backgrounds (things we know) and if there is evidence in the observations for a new phenomenon. This means one needs to test two hypotheses: a background only ( $b$ ) and a signal plus background ( $s + b$ ) hypothesis. Each will result in a p-value on their own.

For example  $p_b = 0$  would mean that the backgrounds are perfectly reflected by the observations and a  $p_{s+b} < 0.05$  could be a sign of e.g. new physics. To combine these two metrics into a single score, particle physicists came up with the pseudo Confidence Level/p-value called  $\text{CL}_s$  incorporating also the goodness of the modeling of the backgrounds

$$\text{CL}_s = \frac{p_{s+b}}{1 - p_b} = \frac{\int_{t_{\mu,obs}}^{\infty} f(t_{\mu} | \mu) dt_{\mu}}{1 - \int_{t_{\mu,obs}}^{\infty} f(t_{\mu} | \mu) dt_{\mu}}. \quad (1.4.11)$$

Intuitively the numerator is again just the value for the alternative hypothesis whereas the denominator penalizes  $\text{CL}_s$  if the modeling of the backgrounds is not reflected in the observations. This can also be understood visually from the first figure of the heavily cited  $\text{CL}_s$  paper [8] (see description of fig. 1.2).

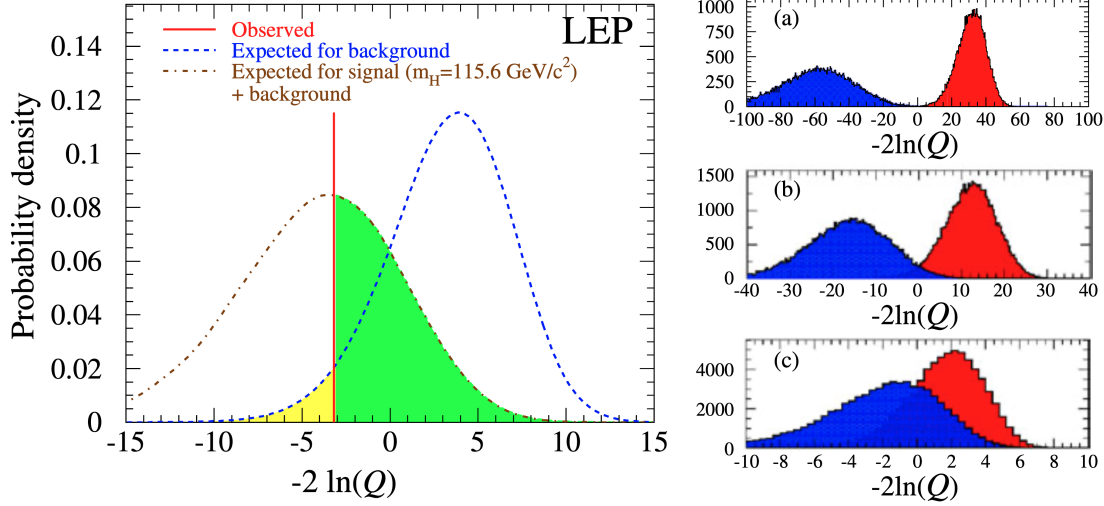
#### 1.4.4 The HistFactory model

A model used widely to achieve the preceding for a typical ATLAS analysis is called HistFactory [9] and is implemented within PYHF ???. This follows closely the introduction to pyhf ???. HistFactory draws the name from the idea that the problem is reducible to a small number of basic components. In order to build a likelihood like eq. 1.4.4 it is again useful to think of another splitting of the nuisance parameters  $\phi$  into

$$L(\mathbf{x}|\phi) = \underbrace{L(\mathbf{x}|\underbrace{\psi}_{\text{nuisance parameters}}, \underbrace{\theta}_{\text{parameters of interest}})}_{\text{nuisance parameters}} = L(\mathbf{x}|\underbrace{\eta}_{\text{free}}, \underbrace{\chi}_{\text{constrained}}) \quad (1.4.12)$$

free parameters  $\eta$ , like a cross section of a process and constrained parameters  $\chi$ , to account for systematic uncertainties. The likelihood can then be modeled with Poissonians for each of the channels  $c$  (e.g. signal regions), bins  $b$

$$L(\mathbf{n}, \mathbf{a} | \eta, \chi) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}(n_{cb} | \nu_{cb}(\eta, \chi))}_{\text{Simultaneous measurement of multiple channels}} \underbrace{\prod_{\chi \in \chi} c_{\chi}(a_{\chi} | \chi)}_{\text{constraint terms for auxiliary measurements}}. \quad (1.4.13)$$



**Figure 1.2:** Probability density functions of test statistics from a Higgs search at LEP illustrating the calculation of p-values ( $\lambda$  becomes  $Q$ ). **(left)** The pdf's of the test statistic  $f(t_\mu | \mu)$  of the signal + background ( $\text{---}$ ) and background ( $\text{---}$ ) only hypotheses. The p-value is calculated by integration from  $t_{\mu,obs}$  (the red observed line ( $\text{---}$ )) to infinity (see eq. 1.4.7). The green shaded area ( $\text{■}$ ) corresponds to  $p_{s+b}$  whereas the yellow area ( $\text{■}$ ) corresponds to  $1 - p_b$  since the integral over one whole pdf is 1. **(right)** Degradation of search sensitivity from (a) to (c). Note that the colors of the pdf's change here to signal + background ( $\text{■}$ ) and background only ( $\text{■}$ ). For example putting the observation ( $t_{\mu,obs}$ ) on the x-axis at 0 in these plots, one would get for plot (a)  $p_b \approx 1$  and  $p_{s+b} \approx 0$  resulting in a  $CL_s \approx 0$ , whereas with increasing overlap the  $CL_s$  value increases and the sensitivity decreases. From [8].

The  $n_{cb}$  is the observation itself and  $\nu_{cb}(\boldsymbol{\eta}, \boldsymbol{\chi})$  the prediction. The  $c_\chi$  are constraint terms calculated from auxiliary measurements  $a_\chi$  to constrain  $\chi$  and can be any function (e.g. Gaussian, Poissonian,...) we believe the parameter to be distributed.

The prediction is a sum over all samples (e.g.  $t\bar{t}$ , multijet-background etc.) which are modified by multiplicative and additive modifiers

$$\nu_{cb}(\phi) = \sum_{s \in \text{samples}} \nu_{scb}(\boldsymbol{\eta}, \boldsymbol{\chi}) \quad (1.4.14)$$

$$= \sum_{s \in \text{samples}} \underbrace{\left( \prod_{\kappa \in \boldsymbol{\kappa}} \kappa_{scb}(\boldsymbol{\eta}, \boldsymbol{\chi}) \right)}_{\text{multiplicative modifiers}} \left( \nu_{scb}^0(\boldsymbol{\eta}, \boldsymbol{\chi}) + \underbrace{\sum_{\Delta \in \boldsymbol{\Delta}} \Delta_{scb}(\boldsymbol{\eta}, \boldsymbol{\chi})}_{\text{additive modifiers}} \right). \quad (1.4.15)$$

$$\text{Var}(\text{Pois}(k, \lambda)) = \sigma^2 = \text{lambda} = \delta^2 \rightarrow \lambda = \sqrt{\delta} \quad (1.4.16)$$

$$\text{Pois}(k, \frac{1}{\sqrt{\delta}}) \quad (1.4.17)$$

# Bibliography

- [1] A. Zee, *Quantum field theory in a nutshell*, Vol. 7 (Princeton university press, 2010).
- [2] D. Griffiths, *Introduction to elementary particles* (John Wiley & Sons, 2020).
- [3] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, The European Physical Journal C **71**, 1 (2011).
- [4] O. Behnke, K. Kröninger, G. Schott, and T. Schörner-Sadenius, *Data analysis in high energy physics: a practical guide to statistical methods* (John Wiley & Sons, 2013).
- [5] L. Heinrich, “Introduction to model building within pyhf,” <https://pyhf.readthedocs.io/en/v0.7.2/intro.html> (2021), [Online; accessed 27-June-2023].
- [6] A. Wald, Transactions of the American Mathematical society **54**, 426 (1943).
- [7] S. S. Wilks, The annals of mathematical statistics **9**, 60 (1938).
- [8] A. L. Read, Journal of Physics G: Nuclear and Particle Physics **28**, 2693 (2002).
- [9] K. Cranmer, A. Shibata, W. Verkerke, L. Moneta, and G. Lewis, “Histfactory: A tool for creating statistical models for use with roofit and roostats,” Tech. Rep. (2012).



### **Statutory Declaration - Eidesstattliche Erklärung**

I declare that I have authored this thesis independently, that I have not used other than the declared sources/ resources and that I have explicitly marked all materials which has been quoted either literally or by content form the used sources.

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Berlin, 30.06.2023

---

Frederic Renner