

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, 28.06.2023

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

1	Theory	1
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	6
1.4.4	Histfactory in pyhf	7
	Bibliography	9

List of Figures

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 σ . 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 δ^4 ensures energy and momentum conservation, the last δ 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 Θ 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 \rightarrow qed Lagrangian

gauge field blah only about Lagrangians, no field solutions needed

1.4 Statistics

Every science starts from a hypothesis to be tested. The task is to make a statement 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.

For exactly this task high energy physicists have developed a framework based on likelihood statistics tailored for counting experiments. This section begins to lay out the 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

As we are dealing with a counting experiment the tool at hand are histograms $\mathbf{n} = (n_1, \dots, n_N)$. It can be modeled with a set of parameters divided into so called parameters of interest, here only the signal strength μ , and nuisance parameters Θ , that basically serve to give the model flexibility to fit the observations. 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. The expectation value of the n_i is then

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

The model can be further constrained with auxiliary histograms $\mathbf{a} = (a_1, \dots, a_M)$ with bin height

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

As we are expecting the bin counts to occur with a constant mean rate and independent of time compared to the last event each 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. From this a likelihood $L(\mathbf{x})$ can be built, which is just a probability under a given set of parameters \mathbf{x} . Accounting for all

the bins by multiplying them together yields

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

To test for a hypothesized value of μ we consider the so called profile likelihood ratio that reduces the dependence to one parameter of interest μ

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

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 μ and the set of found nuisance parameters $\hat{\Theta}$ that maximize the likelihood. This definition gives $0 \leq \lambda \leq 1$. For a $\lambda \sim 1$ the hypothesized value of μ shows good agreement to the Poissonian model.

1.4.2 From test statistic to p-value

Transforming this into a test statistic t_μ 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 p-value can then be calculated from the probability density function of t_μ : $\text{pdf}(t_\mu) = f(t_\mu | \mu)$

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

$t_{\mu,obs}$ is the test statistic t_μ evaluated at the observed data. This is like plugging into the Poisson distributions the same values for r as for k in eq. 1.4.3¹. Just like a probability density function for a standard normal distribution, intuitively the pdf is just how probable the test statistic t_μ is under a fixed value of the signal strength (how often it occurs compared to all other values t_μ can be). This particular form is handy because there exist approximations for $f(t_\mu | \mu)$ (for details see [3]). Figure

¹The subindex μ for the observation in $t_{\mu,obs}$ can be a bit misleading because plugging in actually drops the dependency on μ .

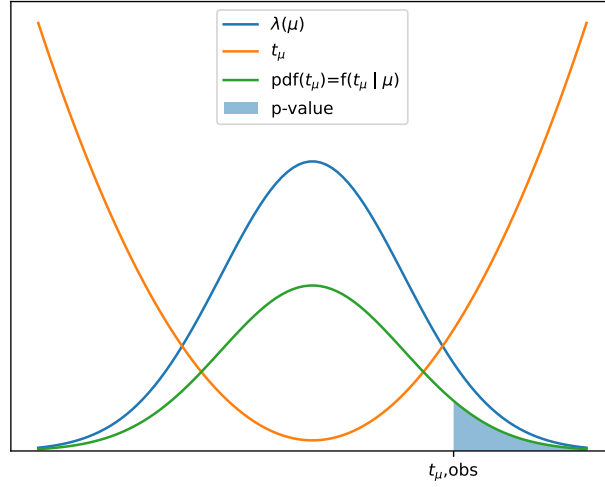


Figure 1.1: This is a sketch to mentally trace the transformation steps. The profile likelihood has essentially a hill like form with a maximum at $\lambda(\hat{\mu}, \hat{\Theta})$, t_μ is $\sim -\ln(\lambda)$ and under the assumption that the Poissonian model is correct the pdf of t_μ should peak at $\lambda(\hat{\mu}, \hat{\Theta})$

1.1 illustrates the different steps as a sketch. One caveat here is that this assumes μ can also take negative values which of course is not the intended goal here. But considering the correct limiting cases one can construct test statistics accordingly ([3] for details). In the scientific community a widely accepted threshold to accept an alternative hypothesis over the null hypothesis is for a p-value of 0.05.

1.4.3 The CL_s value

Being able to calculate p-values allows now to state how likely it is that the proposed hypothesis is reflected by the observed data. Taking the example from the sketch in fig. 1.1 one sees that the blue shaded area makes up about $\sim 10\%$ of the total area below the curve of $f(t_\mu | \mu)$. Since $f(t_\mu | \mu)$ is a pdf it holds that $\int f(t_\mu | \mu) = 1$ and therefore gives a p-value of $p \sim 0.1$. With other words it means if we were to repeat the experiment, in $\sim 10\%$ of the cases the observed data would be in agreement with our proposed hypothesis.

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 can we find 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 $p_b = 1$ would mean that the backgrounds are perfectly reflected by the observations and a $p_{s+b} < 0.05$ could be a sign of new physics. To combine these two metrics into a single score particle physicists came up with the pseudo Confidence Level called CL_s incorporating also the goodness of the modeling of the backgrounds

$$CL_s = \frac{CL_{s+b}}{CL_b} = \frac{\int_{t_{\mu=1,obs}}^{\infty} f(t_{\mu} | \mu = 1) dt_{\mu}}{\int_{t_{\mu=0,obs}}^{\infty} f(t_{\mu} | \mu = 0) dt_{\mu}}. \quad (1.4.8)$$

This can also be understood visually from the first figure of the heavily cited CL_s paper [6] (fig. 1.2). The left figure depicts the pdf's of the test statistic $f(t_{\mu} | \mu = 0)$ of the signal + background (orange) and background (blue) only hypotheses. The p-value is calculated by integration from $t_{\mu,obs}$ (the red observed line) to infinity (see eq. 1.4.7). The green shaded area corresponds to p_{s+b} whereas the integral from the red line to infinity for the blue curve would be p_b .

The right hand side depicts the degradation of search sensitivity (a)→(c). Unfortunately the colors of the pdf's changed here to signal + background (blue) and background only (red). For example if you put the observation on the x-axis at 0 in these plots, you would get for the first one basically $p_b \approx 1$ 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 decrease.

1.4.4 Histfactory in pyhf

Detector-simulation related uncertainty – Calibrations (electron, jet energy scale) – Efficiencies (particle ID, reconstruction) – Resolutions (jet energy, muon momentum)! ! • Theoretical uncertainties – Factorization/Normalization scale of MC generators – Choice of MC generator (ME and/or PS, e.g. Herwig vs Pythia) • Monte Carlo Statistical uncertainties – Statistical uncertainty of simulated samples

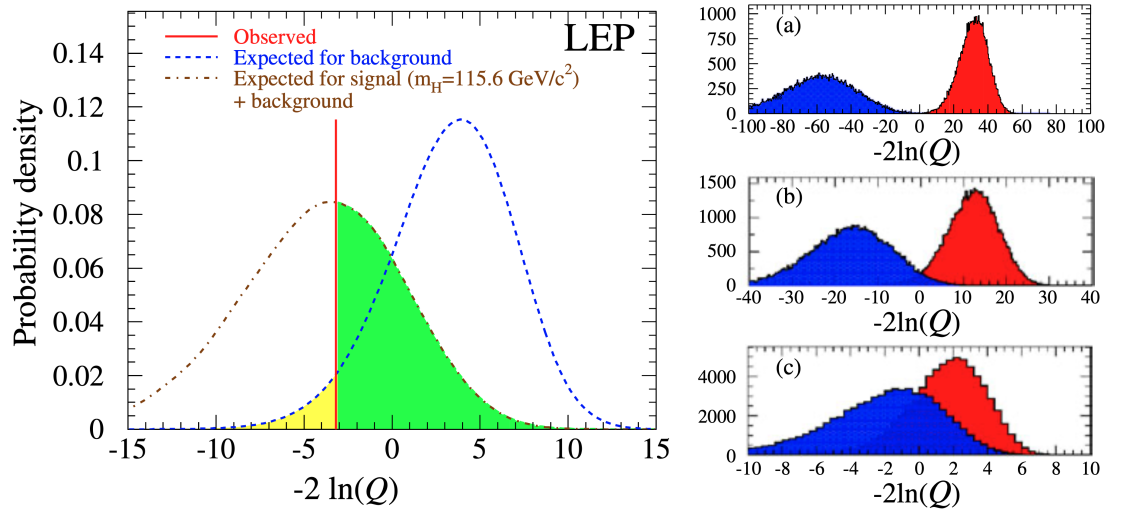


Figure 1.2: Probability density functions of test statistics from a Higgs search at LEP illustrating the calculation of p-values. Description at the end of section 1.4.3. From [6].

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. L. Read, Journal of Physics G: Nuclear and Particle Physics **28**, 2693 (2002).

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, 28.06.2023

Frederic Renner