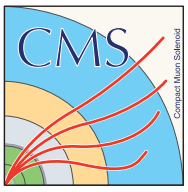
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Optimization of classifier binning for differential statistical inference in particle physics

Master Thesis

angefertigt im
Institut für Experimentalphysik
vorgelegt der
Fakultät für Mathematik, Informatik und Naturwissenschaften



Jan Hauke Voß

Born: 25.10.1997
Master of Science Physik
Matr.-Nr. 6925999

Erstgutachter: Professor Dr. Peter Schleper
Zweitgutachter: Professor Dr. Johannes Haller

Hamburg, den 05.08.2023

Jan Hauke Voß

Optimization of Classifier Binning for Differential Statistical Inference in Particle Physics

Masterarbeit in Physik

Universität Hamburg

Institut für Experimentalphysik

Erstgutachter: Prof. Dr. Peter Schleper
Zweitgutachter: Prof. Dr. Johannes Haller

Zusammenfassung

Statistische Inferenz spielt in der Teilchenphysik eine wichtige Rolle. In dieser Arbeit werden neue Methoden entwickelt, um zusammenfassende Statistiken durch optimiertes Binning von Klassifikatoren zu optimieren. Es werden zwei Methoden vorgestellt, die die erwartete obere 95% $\mathrm{CL_s}$ Konfidenzgrenze für einen Signalstärkeparameter direkt auf dem Diskriminatorausgang oder auf einer analytischen Darstellung davon optimieren. Beide Methoden sind in der Lage, das Binning und den Grenzwert in einem angemessenen Umfang zu optimieren. Die direkte Methode ist jedoch durch statistische Schwankungen begrenzt. Die Methode mit der analytischen Darstellung des Klassifikators kann mit den Schwankungen umgehen, aber die analytische Darstellung muss sorgfältig gewählt werden.

Abstract

Statistical inference plays a key role in particle physics. In this thesis new methods are developed to optimize summary statistics through optimized binning of classifiers. Two methods are presented that optimize the expected upper 95% $\mathrm{CL_s}$ limit for a signal strength parameter on the discriminator output directly or on an analytic representation of it. Both methods are able to optimize the binning and the limit to a reasonable extent. However, the direct method is limited by statistical fluctuations. The method with the analytic representation of the classifier can deal with the fluctuations, but the analytic representation has to be chosen carefully.

Contents

1 Introduction

The sole purpose of particle physics is to unravel the fundamental structures and mechanisms of nature.
To that end models and theories describing these structures and mechanisms are developed and then tested against in experiments.
Currently the most important theory in that regard is the Standard Model of Particle Physics (SM), describing the three fundamental forces that describe all known elementary particles to high levels of precision.
However, the testing of the SM is an ongoing process and there are still predictions within this framework that need to be validated or possibly excluded through experimental evidence.

One of those predictions are the self-couplings of the Higgs boson, involving up to four Higgs bosons sharing an interaction vertex. As the production of Higgs bosons is not a simple task, and the relatively small amount that can be produced today is limiting the search for these self-couplings and, in the worst case, may not be feasible with current experiments. Thus, it is even more important for particle physics to try to improve the analysis methods to gain as much knowledge as possible from the collected data.
In this thesis, novel methods are examined to optimize summary statistics and thereby enhance the statistical inference of analyses.
For this purpose, new tools that enable working with statistics in a differentiable way, are used to develop a new algorithm that automatically optimizes the binning of the distribution utilized for the signal extraction or computation of upper limits. This method should for the first time allow binning "free of assumptions", compared to currently used optimization methods e.g. [11].
The structure of this thesis is as follows: Section 2 gives an introduction to the SM, the Higgs boson and its self-couplings. In Section 3 a few details on the Large Hadron Collider and the Compact Muon Solenoid are presented whose simulation data is used in this thesis. Section 4 introduces the basic statistical core principles and gives an overview over the statistical software and tools used. In Section 5 parts of the di-Higgs analysis flow and methodology are described and a motivation for this thesis is given. Subsequently, Section 6 describes the data pre-processing and new methodology for the binning optimization. Section 7 presents the results of the new methods and Section 8 concludes with a summary and outlook.

2 The Standard Model of Particle Physics

This chapter examines the fundamental content of the Standard Model, explores the properties of the Higgs boson, and introduce the search for Higgs self-couplings.

2.1 Introduction to the Standard Model of Particle Physics

The SM [1] introduces fermions to describe matter, gauge bosons to describe the propagation of the three fundamental forces, and a scalar boson necessary for generating the masses of the particles. The three fundamental forces in question are: electromagnetism, the weak and the strong force. The electromagnetic force is transmitted through the photon that can couple to any electrically charged particles. The weak force is transmitted via the Z and W bosons and couples with particles that have a weak isospin. The strong force is transmitted by the gluon that couples only with color charged particles. The strength of the forces develops differently with distance or energy scale. While the strong force becomes stronger for larger distances, the other two become weaker. For higher energy scales in contrast, the strong and weak forces weaken and the electromagnetic force becomes stronger.
The fields for the three fundamental forces are introduced in the SM theory due to the requirement for the theory to be not only globally but also locally invariant under gauge transformations. These fields give rise to the gauge bosons. However, explicit mass terms would violate the required local gauge invariance. As a consequence, none of the particles would acquire any mass in the SM without the Higgs field, its coupling to matter and force-carrying particles, and the accompanying Higgs boson.

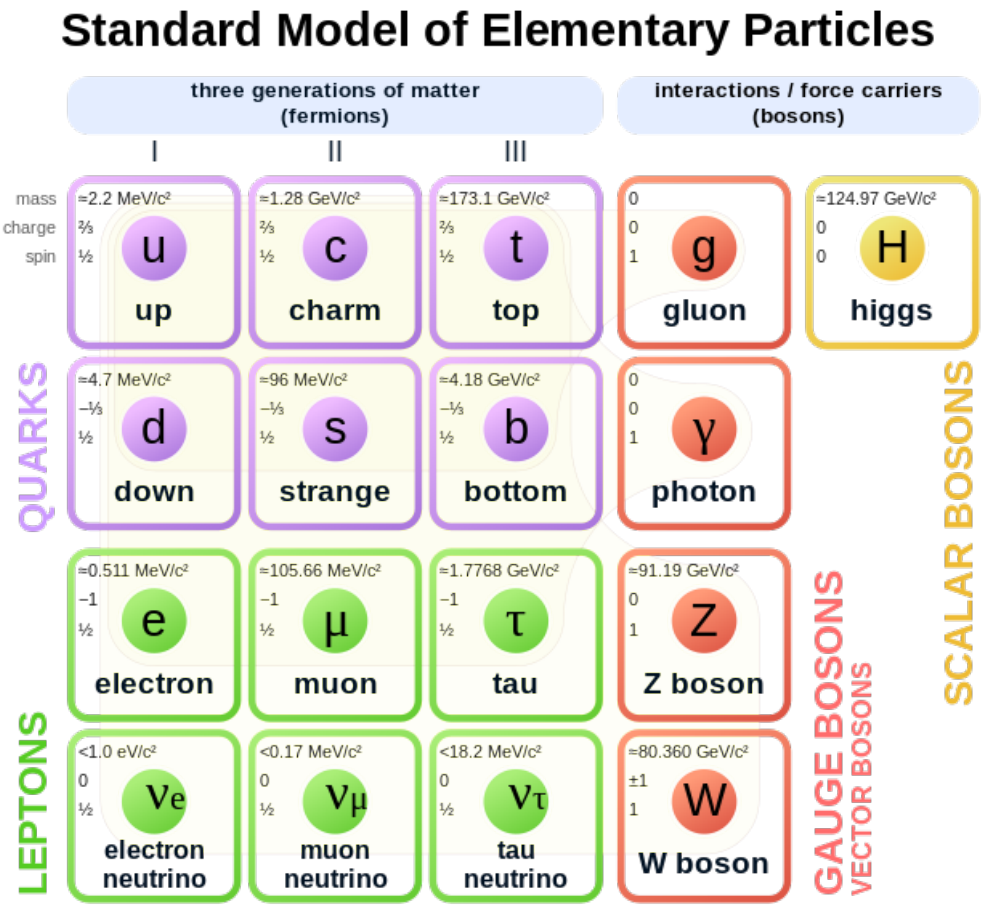


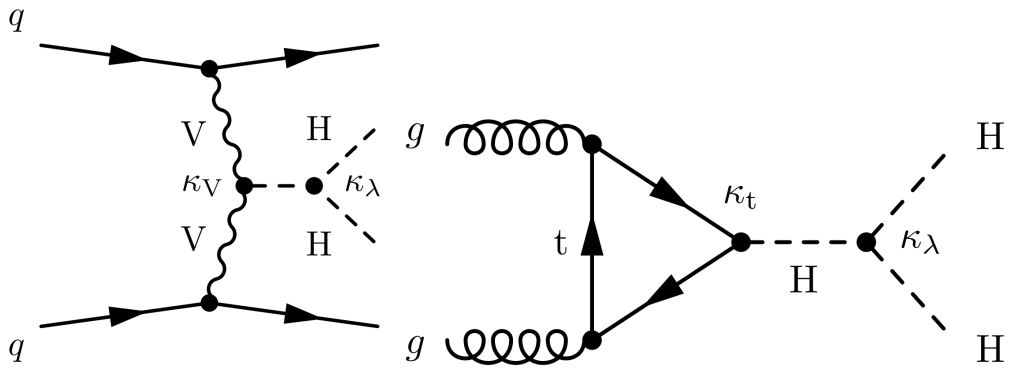
Figure 1: Overview of particles in the SM [17]

2.2 The Higgs Boson and its Self-Couplings

The Higgs mechanism postulates a scalar Higgs field and a gauge-invariant potential for the field. By choosing the free parameter of the potential in a certain way, motivated by renormalization and vacuum stability, the potential acquires a spontaneously broken symmetry and the famous "wine bottle bottom" or "Mexican Hat" shape. Next to other important predictions that follow, such as that the masses of the W and Z bosons are determined by the weak coupling constants and the Higgs potential, the theory also describes Higgs boson self-couplings.
These self-couplings allow to test the shape of the potential, which, as postulated, is an important parameter for the SM and also theories that reach beyond it (BSM).

2.3 The search for Trilinear Higgs Self-Couplings

The main di-Higgs (HH) production channels containing the trilinear self-couplings (λ_{HHH}) can be seen in Figure 2 with the coupling modifier $\kappa_\lambda = \lambda_{\mathrm{HHH}}/\lambda^{\mathrm{SM}}_{\mathrm{HHH}}$. Similarly the coupling modifiers of the Higgs boson to a top quark or a vector boson are labeled $\kappa_{\mathrm{t}}, \kappa_{\mathrm{V}}$.



(a) VBF HH production (b) ggF HH production

Figure 2: HH production channels with trilinear H boson self-coupling at leading order. 2(a) shows the HH production via vector boson fusion (VBF) and 2(b) shows the HH production via gluon-gluon fusion (ggF).

In the SM these production channels have a very small cross section, even when compared to the production of single Higgs bosons, making the detection inherently challenging.

The relative statistical errors scale roughly as $\sigma_{\text{rel,stat}} \sim \frac{1}{\sqrt{N}}$ with the number of recorded data events N. Thus it is essential to use statistical methods that can control these errors within analyses, especially when dealing with small cross sections.

In Figure 3 that is taken from [10] distributions of the cross-sections as function of transverse momentum for Higgs bosons from the di-Higgs process are shown for various values of κ_λ. The SM scenario is $\kappa_\lambda = 1$. For transverse momenta of up to 200 GeV the different predictions for the cross-sections vary strongly.

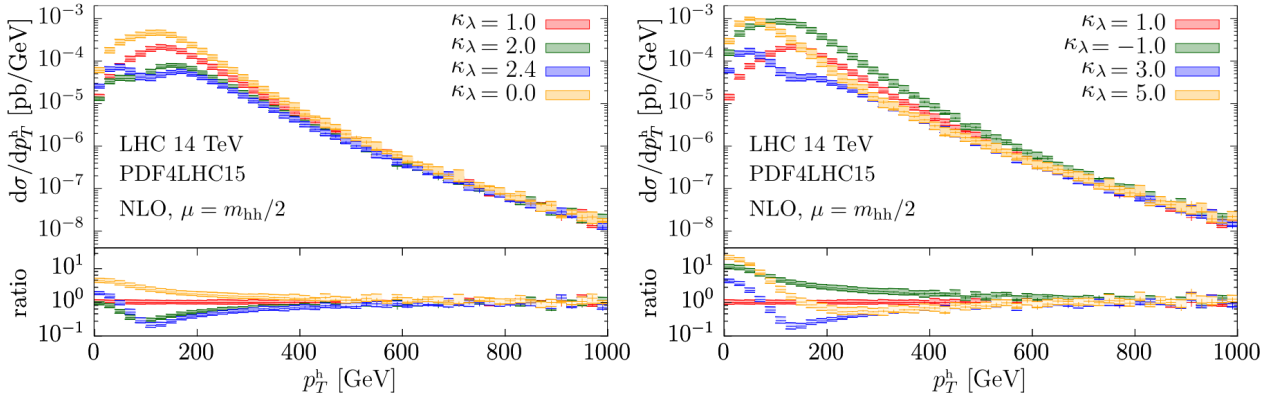


Figure 3: Higgs boson transverse momentum distributions at 14 TeV for the considered κ_λ values [10].

Figure 4, also from [10], shows a 3-D visualisation of the cross-section as a function of the invariant mass of the di-Higgs boson system as function of κ_λ and m_{HH}. In the regions around approximately 300 GeV large variations in the cross-section are present.

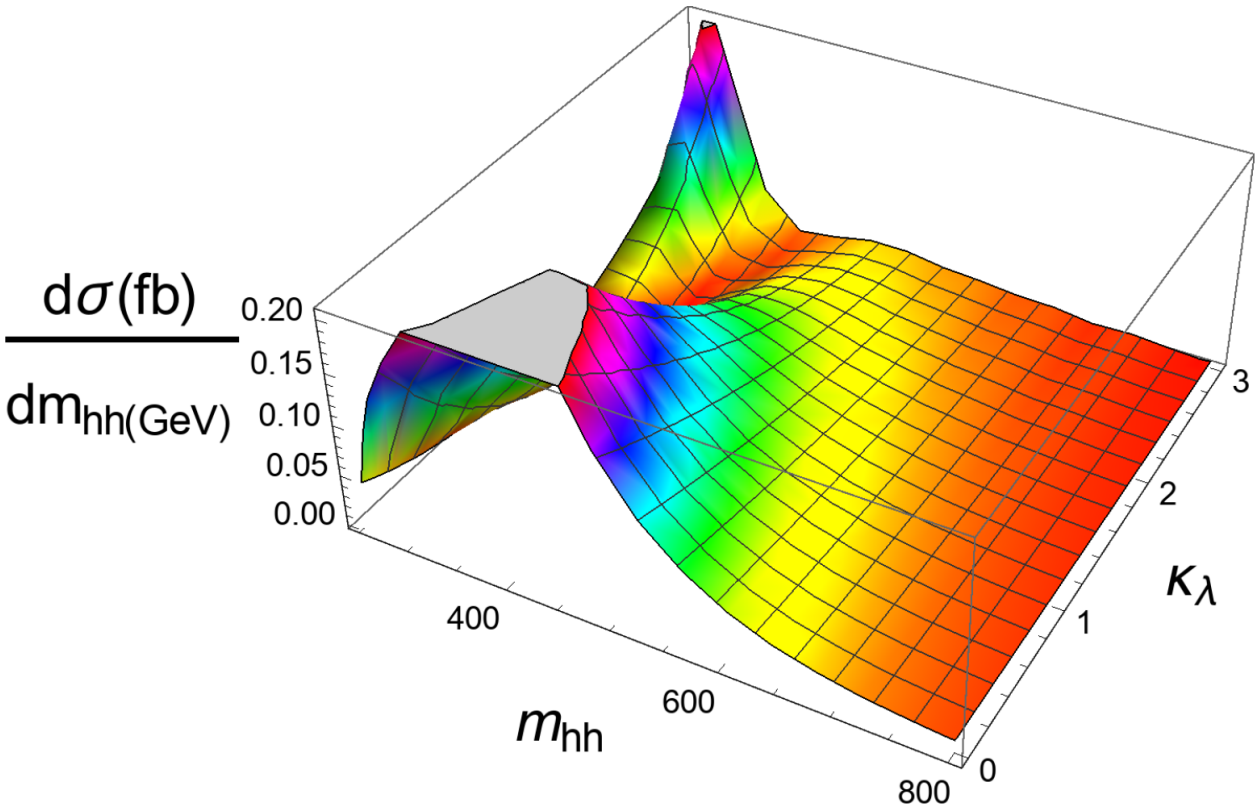


Figure 4: 3-dimensional visualisation of the m_{HH} cross-section at 14 TeV, as a function of κ_λ and m_{HH} [10].

3 Experimental Setup

The data used for this thesis is simulated for the Compact Muon Solenoid detector (CMS) at the Large Hadron Collider (LHC), which is located at CERN. This chapter provides further details about the experiment and the data.

3.1 The LHC and the CMS Detector

The LHC is a circular hadron collider equipped with four large specialized detectors, shown in Figure 5.
LHCb (Large Hadron Collider beauty) is a detector specialized on physics containing b-quarks, ATLAS is a general-purpose detector and ALICE (A Large Ion Collider Experiment) is intended to study heavy-ion collisions.

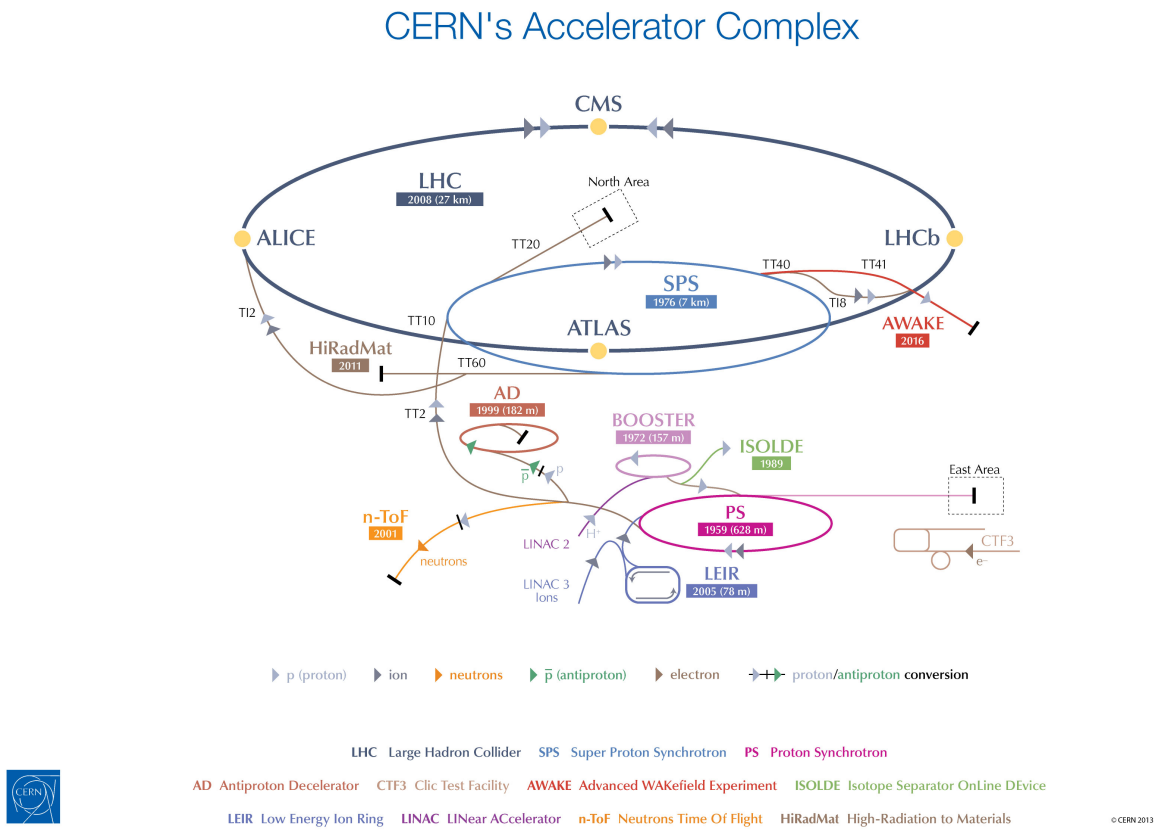


Figure 5: Schematic overview of the LHC and the CERN accelerator complex [9].

The CMS detector, as seen in Figure 6, is a general-purpose detector. It has a barrel shape surrounding the collision point in the beampipe. The innermost part is the tracker system made of silicon pixel detectors and strip detectors. In these components charged particles leave hits, that can be reconstructed into tracks. Around the tracker system is the electromagnetic calorimeter (ECAL) consisting of lead tungstate crystals ($PbWO_4$). In this part of the detector, charged particles and light deposit their energies. These systems are furthermore surrounded by the hadronic calorimeter (HCAL). The HCAL is constructed with alternating layers of brass absorber and plastic scintillators, designed to measure the energies of hadrons as they deposit their energy. Surrounding the previously mentioned layers is a superconducting solenoid, which is partly responsible for the name

CMS. It is made of a niobium titanium coil (NbTi) and can create a magnetic field of up to 4 T inside. This magnetic field is used to curve tracks of charged particles, allowing for the measurement of momentum.
The outermost layer is the name giving muon system. It consists of the muon chambers to detect muons and an iron return yoke for the magnetic field that helps with the momentum measurements of the muons.

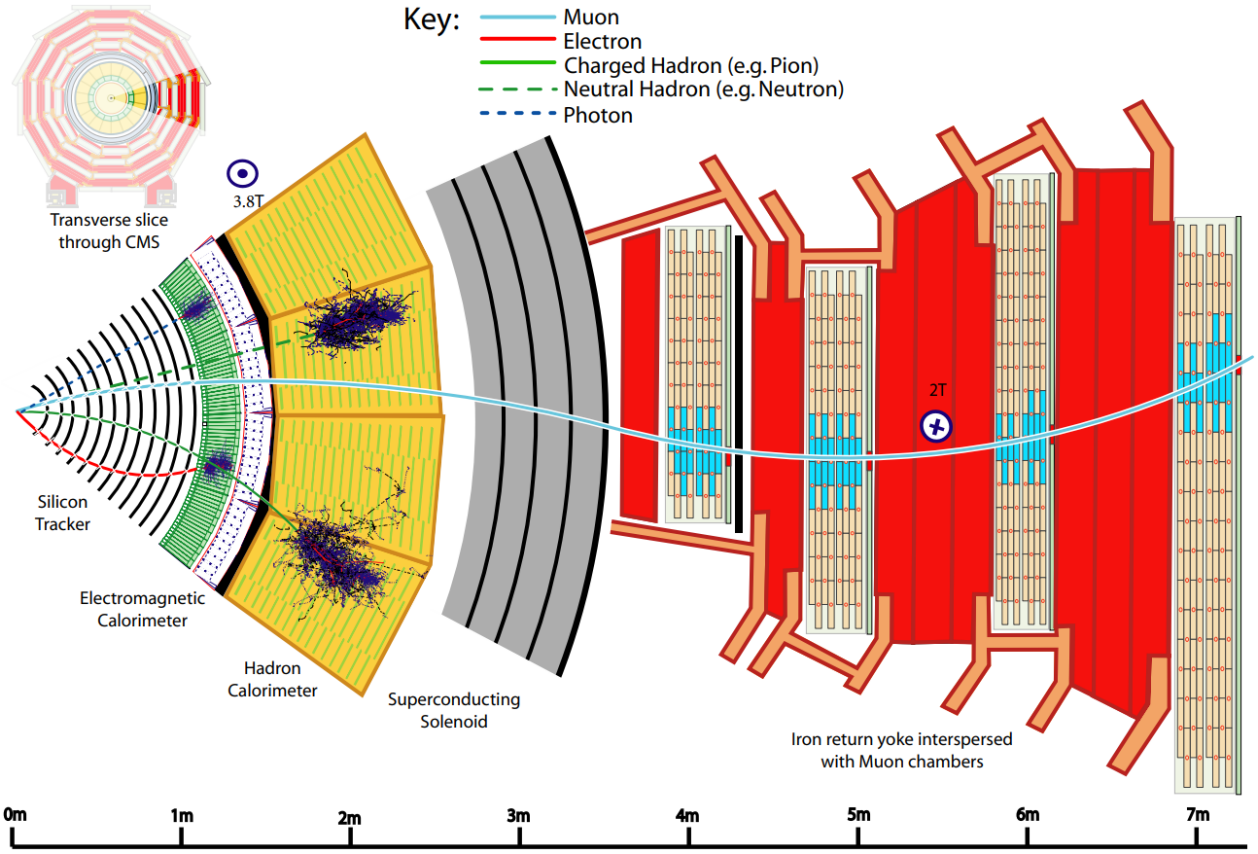


Figure 6: A schematic slice of the CMS detector [14].

3.2 Simulated data

To test theories against the real-world data measured with the detector, data is produced through simulation to represent those theories. This production is achieved with Monte-Carlo (MC) simulations. Events are generated from the theory parameters first and then the detector properties are simulated. Afterwards, the events are processed with the same reconstruction algorithms as the real data. Further weights and corrections are applied afterwards to account for any remaining discrepancies between the real and MC data.
This thesis uses simulated data events selected and processed by the search for HH production using the CMS experiment with the dataset of the second run of the LHC. More details on the analysis can be found in section 5.1 and in the analysis note [2] that also contains a complete description of how the used data was simulated.

4 Statistics in Particle Physics Experiments

This section describes how the upper limit using the $\mathrm{CL_s}$ method is constructed and how it may be interpreted. Also it presents the tools used in this thesis for statistical computations.

4.1 Statistical Inference

When a measurement is performed in particle physics an important question is always how significant the measured data is and if it is compatible with a theoretical model. The amount of counted events in an experiment can be split up into the amount of background events (b) and the amount of signal events (s). To scale the amount of signal for different model scenarios, usually the signal strength modifier μ is introduced. Then the total amount of expected events (ν) is

$$\nu \quad = \quad \mu \cdot \mathrm{s} + \mathrm{b} \tag{1}$$

and μ is the parameter of interest (poi) for the inference on this statistical model. In counting experiments the probability for x counted events, when expecting the number of events ν, is given by the Poisson distribution (P)

$$\mathsf{P}(\mathrm{x}|\nu) \quad = \quad \frac{\nu^{\mathrm{x}}}{\mathrm{x!}} \cdot \mathrm{e}^{-\nu} \quad . \tag{2}$$

For a binned measurement the combined probability (p) for all bins (N) is considered as a combination of multiple independent counting experiments and is therefore the product of each Poisson distribution in each bin (i)

$$p(\mathrm{x}|\nu) \quad = \quad \prod_{\mathrm{i}}^{\mathrm{N}} \mathsf{P}(\mathrm{x_i}|\nu_\mathrm{i}) \quad . \tag{3}$$

To evaluate the parameter of interest regarding a fixed measurement the likelihood function ($\mathcal{L}$) can be formulated as function of the parameters:

$$\mathcal{L}(\mu|\mathrm{x}) \quad := \quad p(\mathrm{x}|\mu) \quad . \tag{4}$$

For real world measurements the statistical model needs to be adjusted for all kinds of further parameters that represent statistical or systematic uncertainties. These so-called nuisance parameters (θ) are estimated theoretically or with experiments and can be incorporated into the model [5]. For the number of, for example, expected background events in a bin the nuisances are implemented as a composition of the nominal background yields and a prior probability distribution function $\pi_{\mathrm{n,p,i}}(\theta)$ of a nuisance (n) evaluated at θ_n and for processes (p):

$$\mathrm{b_i}(\theta) \quad = \quad \sum_{\mathrm{p}} \mathrm{b_{p,i}} \cdot \prod_{\mathrm{n}}^{\dim(\theta)} \pi_{\mathrm{n,p,i}}(\theta) \quad . \tag{5}$$

This is done analogously for the expected signal events. The influence of these nuisances may vary between bins and/or processes and the likelihood expressed with the nuisances is then written as:

$$\mathcal{L}(\mu, \theta|\mathrm{x}) \quad = \quad \prod_{\mathrm{i}}^{\mathrm{N}} \mathsf{P}(\mathrm{x_i}|\mu \cdot \mathrm{s_i}(\theta) + \mathrm{b_i}(\theta)) \quad . \tag{6}$$

Because μ is not restricted, three possibilities have to be considered: positive values, negative values and zero. A value of zero expects no signal and therefore supports the background-only hypothesis. A positive value postulates the presence of signal and is named the signal hypothesis. Negative values for μ have no physical meaning and suggest a statistical deficit of signal-like events in a measurement. For solid inference about the poi one has to differentiate between the two hypotheses and if possible handle the negative values. To achieve that, the likelihood can be modified into a profile likelihood ratio (λ) [6].

$$\lambda(\mu|\mathrm{x}) \quad = \quad \frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu)|\mathrm{x})}{\mathcal{L}(\hat{\mu}, \hat{\theta}|\mathrm{x})} \tag{7}$$

Following from the Neyman-Pearson lemma [12], the likelihood ratio is the most powerful test statistic to evaluate two opposite hypotheses and the profiling allows models with multiple free parameters to make use of it, while also reducing the functions dependencies to only μ. The $\hat{\mu}$ and $\hat{\theta}$ represent the maximum likelihood estimators (MLE) and the $\hat{\hat{\theta}}(\mu)$ represents the conditional MLE. While the MLE maximize the likelihood globally, the conditional MLE maximizes $\mathcal{L}$ for a specific poi value. λ can be further modified to indirectly require a non-negative poi with:

$$\tilde{\lambda}(\mu|\mathrm{x}) \quad = \quad \begin{cases} \frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu)|\mathrm{x})}{\mathcal{L}(\hat{\mu}, \hat{\theta}|\mathrm{x})}, & \hat{\mu} \geq 0 \\ \frac{\mathcal{L}(\mu, \hat{\hat{\theta}}(\mu)|\mathrm{x})}{\mathcal{L}(0, \hat{\hat{\theta}}(0)|\mathrm{x})}, & \hat{\mu} < 0 \end{cases} . \tag{8}$$

To perform statistical tests with this test statistic, knowledge about its probability distribution f is required. Due to Wilks' theorem [18] and generalizations by Wald [16], it can be shown that the transformation of $\tilde{\lambda}$ in Equation 9 approaches a chi-squared distribution with one degree of freedom [6].

$$\tilde{\mathrm{q}}(\mu|\mathrm{x}) \quad = \quad -2\ln\tilde{\lambda}(\mu|\mathrm{x}) \tag{9}$$

This test statistic can be adjusted further by a case distinction to test if a signal process production has a specified rate μ or some rate below that value:

$$\tilde{\mathrm{q}}_{\mu} \quad = \quad \begin{cases} -2\ln\tilde{\lambda}(\mu|\mathrm{x}), & \hat{\mu} \leq \mu \\ 0, & \hat{\mu} > \mu \end{cases} . \tag{10}$$

Using this test statistic and the values $\tilde{\mathrm{q}}_{\mathrm{obs}}$ it yields for measurements x, the significance of x can be calculated regarding different poi and thus different hypotheses. Equations 11 and 12 show the calculations for the p-values regarding compatibility with signal strengths (signal hypotheses) and no signal strength (background-only hypothesis). The necessary distribution $\mathrm{f}(\tilde{\mathrm{q}}_{\mu}|\mu)$ can be determined by using toys or an Asimov dataset [6].

$$\mathrm{CL}_{\mathrm{s+b}} \quad = \quad \int_{\tilde{\mathrm{q}}_{\mathrm{obs}}}^{\infty} \mathrm{f}(\tilde{\mathrm{q}}_{\mu}|\mu)\,\mathrm{d}\tilde{\mathrm{q}}_{\mu} \tag{11}$$

$$CL_b = \int_{\tilde{q}_{obs}}^{\infty} f(\tilde{q}_\mu|0)\,d\tilde{q}_\mu \tag{12}$$

To calculate the frequentist upper limit on a μ at 95% confidence level $CL_{s+b} = 0.05$ has to be solved for μ.
In particle physics this upper limit is modified to protect against statistical downward fluctuations by dividing by the significance for the background-only hypothesis, such that small values in the upper limit are able to also exclude $\mu = 0$:

$$CL_s = \frac{CL_{s+b}}{CL_b} \quad . \tag{13}$$

The construction of this modified p-value is called the CL_s method. Limits or upper limits in this thesis always refer to the modified p-value/ CL_s.

4.2 Statistical Software and Tools

Upper limits calculated with binned likelihoods are strongly influenced by the binning itself. Even though the same events are used the binning decides how the events enter the model. Optimizing the binning regarding the limit requires sophisticated software tools for the calculation of the limits. These tools are presented in the following section.

The implementation of the ideas and algorithms for the novel methods developed in this thesis makes use of some relatively new tools. The statistical models use `pyhf` [8] and are differentiable, meaning that e.g. the CL_s value can be differentiated with respect to the number of events in each bin.
`pyhf` is a pure Python implementation of the HistFactory [7] made for multi-bin histogram-based analysis.
To calculate hypothesis tests `relaxed` [13] is used. `relaxed` is based on `JAX` and provides differentiable versions of common high-energy physics operations e.g. MLE fits, hypothesis tests or histograms.
`JAX` [3] is a software package that combines `Autograd` and `XLA` packages. It can perform automatic differentiation of native Python or NumPy functions and is used to differentiate upper limits in this thesis towards the contents of particular bins.
`combine` [6] is a widely used software package for statistical analysis in CMS analyses. In Section 7.1 its upper limit scan is compared to the setup used in this thesis, because it is widely used.

5 HH Analysis Optimization

This thesis builds upon the "Search for Higgs boson pairs produced through gluon and vector boson fusion in the $\mathrm{bb}\tau\tau$ final state with Run-II data" [4]. As already mentioned in the introduction it is the goal to optimize the binning of the final discriminant between signal and background. This chapter gives an overview over the aspects of the HH analysis flow important for this thesis. A full description of the analysis can be found in the dedicated analysis note [[2], [1]].

5.1 HH Analysis

The analysis first imposes new analysis specific trigger requirements for the $\mathrm{H} \rightarrow \tau\tau$ decay products and new VBF $\mathrm{H} \rightarrow \tau_{\mathrm{h}}\tau_{\mathrm{h}}$ triggers.
Then the candidates for $\mathrm{H} \rightarrow \tau\tau$ pair reconstruction are chosen. This is followed by identification and selection of the $\mathrm{H} \rightarrow \mathrm{bb}$ pairs and VBF jets.
The next step is the event categorization. The categories are shown in Figure 7.
First the events are checked for two VBF jet candidates and, if existing, they are put into the VBF category, which is split into further subcategories by kinematic quantities. In the other case the events are categorized as boosted if they have a so-called fatjet composed of at least two subjets that pass the b-tag loose working point, which is a certain DeepFlavour discriminator score threshold. The other events are marked as resolved with either one of the two b-jet candidates passing the medium working point (res1b) or both passing (res2b).
For the signal extraction a Deep Neural Network (DNN) is trained to discriminate signal from expected background events. An output value close to zero indicates a more "background-like" event and closer to one more "signal-like". The variable distribution is then fit to the data.
Except for the statistical QCD errors, the statistical errors of all other background processes are MC-driven. QCD and its errors are fully data-driven. For the binning of the final discriminator an algorithm that minimizes the expected limits as a function of bin edge positions using a so called Bayesian optimization techniques is used. The histograms of DNN outputs with a fine bin width of 0.001 are prepared for each category. Then, in the order from significant to less significant, starting with a single category the binning is optimized, successively adding more categories and freezing the binning of all previous ones. This is repeated for each channel and year.

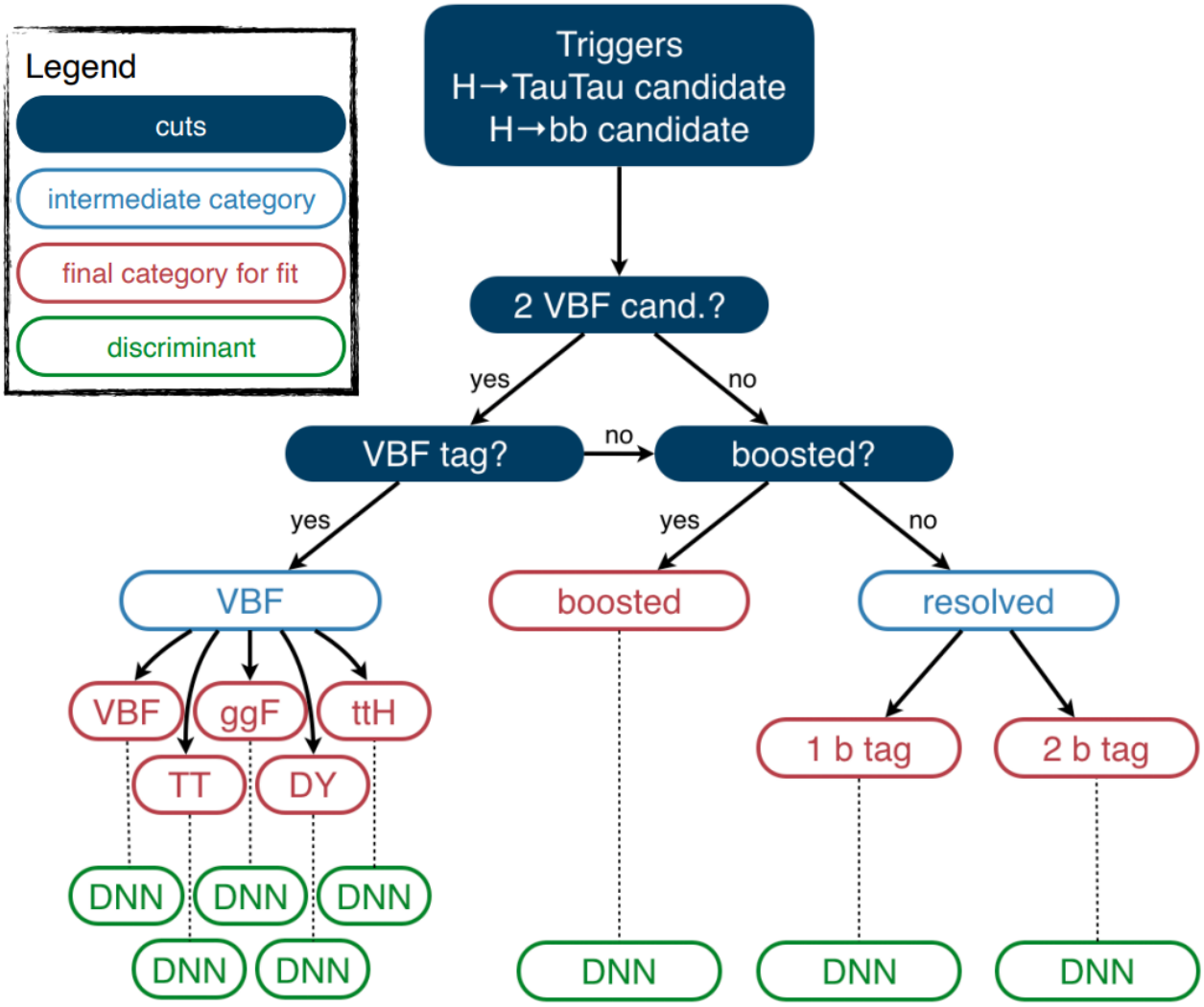


Figure 7: Categories from HH → bb$\tau\tau$ Run 2 analysis. Figure from [1]

The next paragraph describes the shortened process names used in the di-Higgs analysis:

- **QCD**: Data driven background estimation of multi-jet events, that is not described by other MC.
- **DY**: Drell-Yan events.
- **VVV**: Three vector bosons (WWW, WWZ, WZZ and ZZZ).
- **VV**: Two vector bosons (WW, WZ and ZZ)
- **W**: Single W boson.
- **ttH_hbb**: Two top quarks and a Higgs boson. The Higgs boson decays into two b quarks.
- **ggH_htt**: ggF of a Higgs boson that decays into two tau leptons.
- **WH_htt**: A Higgs boson radiated off of a W boson. The Higgs boson decays into two tau leptons.
- **singleT**: A single top or anti-top production in a t-channel.

- **qqH_htt**: VBF production of a Higgs boson that decays into two tau leptons.
- **TT**: A top quark pair.
- **TW**: Single top or anti-top production with a W boson using a b quark.
- **ZH_hbb**: A Higgs boson radiating off of a Z boson and decaying into two b quarks.
- **TTX**: TT with one or two vector bosons (TTW, TTZ, TTWW, TTWZ and TTZZ).
- **EWK**: VBF production of a W or Z boson.
- **gHH_kl_1_kt_1_hbbhtt**: ggF production of di-Higgs, assuming SM couplings.
- **qqHH_CV_1_C2V_1_kl_1_hbbhtt**: VBF production of di-Higgs, assuming SM couplings.
- **QCD**: Data driven background estimation of multi-jet events, that is not described by other MC.
- **DY**: Drell-Yan events.
- **VVV**: Three vector bosons (WWW, WWZ, WZZ and ZZZ).
- **VV**: Two vector bosons (WW, WZ and ZZ)
- **W**: Single W boson.
- **ttH_hbb**: Two top quarks and a Higgs boson. The Higgs boson decays into two b quarks.
- **ggH_htt**: ggF of a Higgs boson that decays into two tau leptons.
- **WH_htt**: A Higgs boson radiated off of a W boson. The Higgs boson decays into two tau leptons.
- **singleT**: A single top or anti-top production in a t-channel.
- **qqH_htt**: VBF production of a Higgs boson that decays into two tau leptons.
- **TT**: A top quark pair.
- **TW**: Single top or anti-top production with a W boson using a b quark.
- **ZH_hbb**: A Higgs boson radiating off of a Z boson and decaying into two b quarks.
- **TTX**: TT with one or two vector bosons (TTW, TTZ, TTWW, TTWZ, TTZZ).
- **EWK**: VBF production of a W or Z boson.
- **gHH_kl_1_kt_1_hbbhtt**: ggF production of di-Higgs, assuming SM couplings.
- **qqHH_CV_1_C2V_1_kl_1_hbbhtt**: VBF production of di-Higgs, assuming SM couplings.

5.2 Optimization Strategy: Motivation

The methodology described in the last section is challenging, as it assumes that training the DNN on discriminating physics motivated classes is optimal for the final inference. This is not necessarily the case and its impact could vary depending on the analysis. Also for the binning optimization the inclusion of one category after another and freezing the previous binnings is done to reduce computational complexity of the brute force method used. Brute force optimizations are not optimal regarding computing resources, and the sequential freezing itself may even block optimal configurations. Furthermore is assumed that the order in which the categories are stacked in is optimal when ordered by sensitivity. It would be optimal to directly train the DNN, and optimize the binning with respect to the final inference without intermediate steps. If done correctly assumptions are not needed and it would give optimizations a maximum amount of information to work with.
This thesis presents a novel methodology and setup for the binning optimization. The methodology is presented in the next section.

5.3 Optimization Strategy: Methods

As discussed in the previous section, the goal is to optimize binning on the final inference. In the case of this thesis it will be the expected upper limit, described in Section 4.1, that is optimized on. For optimization the limit calculation needs to be differentiable, which is made possible with the tools described in Section 4.2. They allow to differentiate the limit with respect to the bin contents and to predict possible changes of bin contents or the binning. To look at smallest possible changes in the binning, the finest binning the DNN can provide is needed. Here the fine bin width of 0.001 is used (1000 bins between 0 and 1) that was also used in the published analysis. Bins from this finest binning will in the following be referred to as mini-bins. Changes in binning can be equivalently described in terms of either mini-bins, bin yields or bin edge positions. The description via yields is only possible if completely empty mini-bins are ignored.
This thesis investigates two different methods for optimization described in Sections 6.4.1 and 6.4.2. One is using the mini-bins directly with a brute force calculation. It depends strongly on event statistics and therefore, is potentially subject to statistical fluctuations.
The other method approximates the measurement distribution analytically and is thus less dependent on the event statistics.
There is a similarity between over-training in machine learning and the introduced methods. The first method can potentially overfit, by including every data point and all fluctuations, which the second method tries to suppress fluctuations by approximating discriminator distributions by spline fits.
Necessary for both methods are a set of constraints discussed in Section 6.4.3. These constraints insure the credibility of the inference they are supposed to optimize, but in return limit the statistical power of the optimization. This trade-off has to be balanced for each analysis by an individual set of constraints, because there is no ansatz based on "first principles" from which to derive them.

6 Optimization Methodology

This section first discusses the motivation for optimization of the signal extraction procedure and describes the novel methods that are explored for the binning optimization. Then it describes the general method and the two different approaches used for the binning optimization, as well as the pre-processing of the data. As pointed out in Section 6.2 the first method is affected strongly by event statistics, which the second tries to bypass. The pre-processing plays a key part, as it allows the usage of statistics tools like pyhf in Section 4.2.

6.1 Optimization Strategy: Motivation

The methodology described in the last section is challenging, as it assumes that training the DNN on discriminating physics motivated classes is optimal for the final inference. This is not necessarily the case and its impact could vary depending on the analysis. Also for the binning optimization the inclusion of one category after another and freezing the previous binnings is done to reduce computational complexity of the brute force method used. Brute force optimizations are not optimal regarding computing resources, and the sequential freezing itself may even block optimal configurations. Furthermore is assumed that the order in which the categories are stacked in is optimal when ordered by sensitivity. It would be optimal to directly train the DNN, and optimize the binning with respect to the final inference without intermediate steps. If done correctly assumptions are not needed and it would give optimizations a maximum amount of information to work with.
This thesis presents a novel methodology and setup for the binning optimization. The methodology is presented in the next section.

6.2 Optimization Strategy: Methods

As discussed in the previous section, the goal is to optimize binning on the final inference. In the case of this thesis it will be the expected upper limit, described in Section 4.1, that is optimized on. For optimization the limit calculation needs to be differentiable, which is made possible with the tools described in Section 4.2. They allow to differentiate the limit with respect to the bin contents and to predict possible changes of bin contents or the binning. To look at smallest possible changes in the binning, the finest binning the DNN can provide is needed. Here the fine bin width of 0.001 is used (1000 bins between 0 and 1) that was also used in the published analysis. Bins from this finest binning will in the following be referred to as mini-bins. Changes in binning can be equivalently described in terms of either mini-bins, bin yields or bin edge positions. The description via yields is only possible if completely empty mini-bins are ignored.
This thesis investigates two different methods for optimization described in Sections 6.4.1 and 6.4.2. One is using the mini-bins directly with a brute force calculation. It depends strongly on event statistics and therefore, is potentially subject to statistical fluctuations.
The other method approximates the measurement distribution analytically and is thus less dependent on the event statistics.
There is a similarity between over-training in machine learning and the introduced methods. The first method can potentially over fit, by including every data point and all fluctuations, which the second method tries to suppress fluctuations by approximating discriminator distributions by spline fits.
Necessary for both methods are a set of constraints discussed in Section 6.4.3. These constraints require a minimum precision on the expected number of background events after a bin optimization. This insures the credibility of the inference they are supposed to optimize, but in return limits the

statistical power of the optimization. This trade-off might have to be balanced for each analysis by an individual set of constraints, as so far there is no ansatz based on "first principles" from which to derive them.

6.3 Data pre-processing

The data set used in this thesis is provided in the form of CMS datacards. The data used is from the hh_res2b_tauTau_2018_13TeV category from the analysis described in Section 5.1, as it is the most sensitive of the categories. In Figure 7 it is the category on the outer right side. It contains a pair of τ lepton candidates as well as two resolved b-jet candidates that both passed the medium working point from 2018 with a center of mass energy of the colliding protons of 13 TeV.

Because this thesis does not use the combine software for statistical inference, for which the datacards are designed, but pyhf, the data has to be converted into a pyhf useable format, in this case the JSON file format.
A comparison and compatibility of the tools is shown in Section 7.1.

The data set contains a certain amount of negative mini-bin yields. These are negative event weights from the generator that arise in the simulation process. For small amounts these can be just set to zero, for large amounts analysis specific solutions need to be found. To test the method in this thesis the negative yields and the corresponding errors are set to zero. Because the statistical errors of the QCD data are not MC but data-driven, this background gives rise to very high relative errors. To simplify the introduction of the method, relative statistical errors are limited to 100% of the expected event yield.
Furthermore the observed data from the datacard is replaced with the sum of all background processes for all limit calculations to arrive at a binning that does not depend on real data.
Due to some technical incompatibilities in the usage of pyhf with JAX, a model with only a single uncertainty can be used for now in the optimizations. As an example of a systematic the three percent uncertainty on the luminosity is selected and applied on each process.

6.4 Binning optimization algorithm

For the limit optimization an analytic expression for changes in the limit regarding binning is needed, connecting the change of a bin content with the respective change of the limit on the signal strength parameter. For simplicity we first consider the problem, for just two bins.

To formulate the change of the limit for changes due to the boundary between two neighbouring bins, it should be noted that the change in yield ($\Delta\mathrm{yield}$) is opposite for the two bins

$$\Delta\mathrm{yield}_1 \quad = \quad -\Delta\mathrm{yield}_2 \quad . \tag{14}$$

The indices refer to two neighbouring bins.
Looking at just the change of signal (s) in one of the bins, the resulting change of limit (L) can be approximated as

$$\Delta\mathrm{L} \quad \approx \quad \Delta\mathrm{s}\left(\frac{\partial\mathrm{L}}{\partial\mathrm{s}}\right) \quad . \tag{15}$$

Using Equations 14, 15 and including the second bin, as well as some background (b), the expression becomes

$$\Delta L_{1,2} \quad \approx \quad \Delta s_1 \left(\frac{\partial L}{\partial s_1} - \frac{\partial L}{\partial s_2} \right) \; + \; \Delta b_1 \left(\frac{\partial L}{\partial b_1} - \frac{\partial L}{\partial b_2} \right) \quad . \tag{16}$$

This formula is now expressed as dependent on only the change in each process and allows for an algorithm to look for the optimal changes. The formula can be generalized for any pairing of neighbouring bins (i, i+1) with any kind of processes (p) inside

$$\Delta L_{i,i+1} \quad \approx \quad \sum_p \Delta p_i \left(\frac{\partial L}{\partial p_i} - \frac{\partial L}{\partial p_{i+1}} \right) \quad . \tag{17}$$

The following two sections show methods to find optimal changes to the binning and the third one discusses how to handle the application of those to the binning.

6.4.1 Mini-bin sorting algorithm

In this approach mini-bins are grouped into a much smaller number of bins. For the later the bin boundaries need to be determined. The direct approach to obtain the best possible changes of bin boundaries using the mini-bins would be to calculate all possible combinations of mini-bins for the change in limit. Doing so for 1000 mini-bins is not sensible. Thus the optimization is done on a given start binning with the number of bins fixed. This limits the combinatorics drastically. The changes in binning can be further constrained by a maximum amount of mini-bin steps, further referred to as the step size. Constraints on the step size or the binning itself are elaborated on in Section 6.4.3.

The sorting algorithm consists of the first three of the following five steps. Steps four and five incorporate it into the whole optimization iteration.

1. Calculate the gradients of the limit via the statistical tools in Section 4.2 for each bin and process.

2. Calculate the ΔL with Equation 17 for each possible bin boundary movement within the maximal step size and other constraints.

3. Determine all moves with $\Delta s, \Delta b$: $\Delta L(\Delta s, \Delta b) < 0$.

4. Apply the stepping algorithm described in Section 6.4.3.

5. Start over until convergence conditions stop the optimization.

An illustration of the whole iteration cycle can be seen in Figure 8.

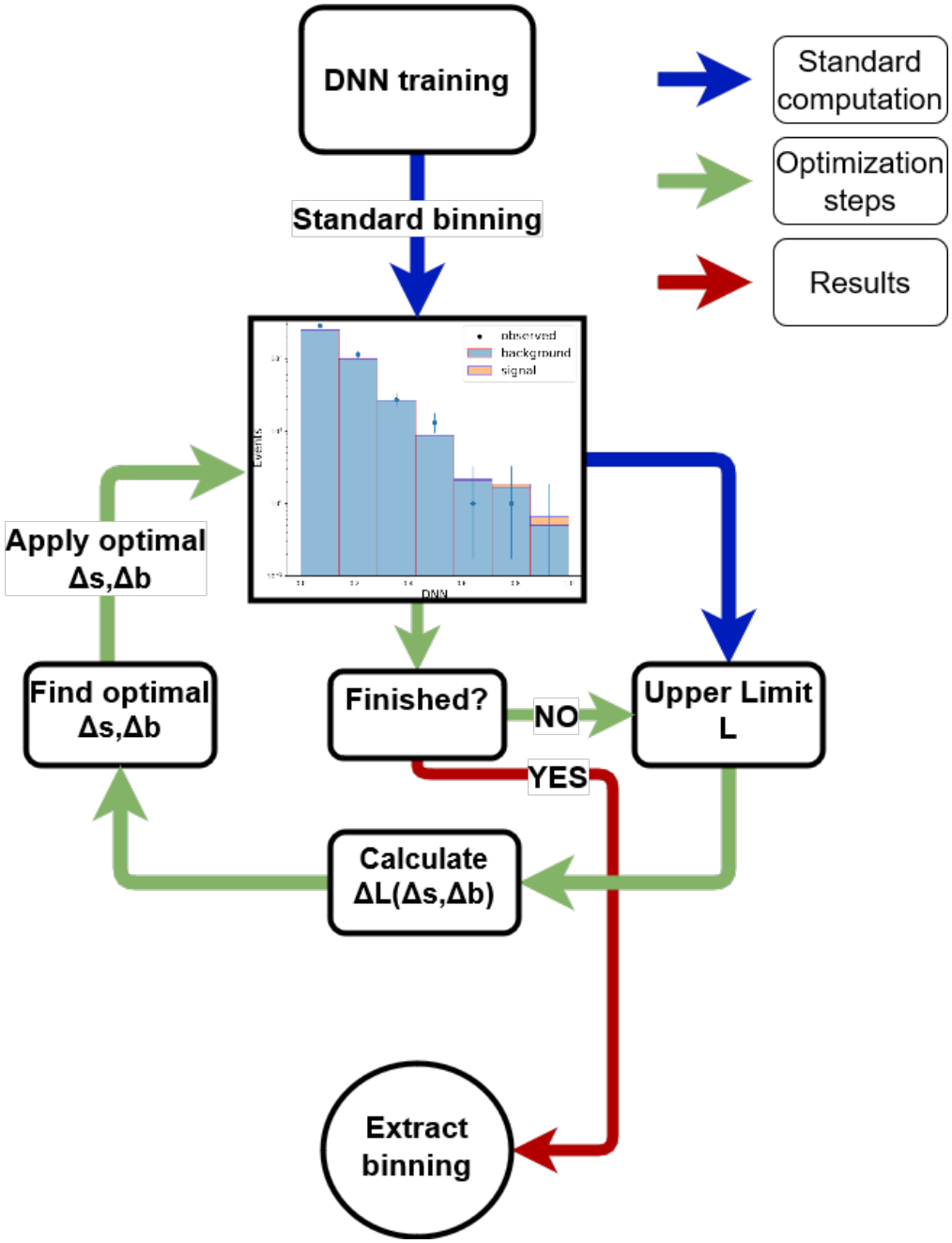


Figure 8: Method overview.

The steps of the algorithm are dependent on the quantity of inference, because it is the goal to improve the upper limit as much as possible. For other quantities the calculation can be adapted accordingly, if needed.
Calculating yields and their errors for new bins is straight forward. The yields are summed up from all the mini-bins a bin contains. The statistical errors on the yields are uncorrelated and therefore the new errors are the square root over the sum of variances (var) of all mini-bins, here mb, contained in a bin

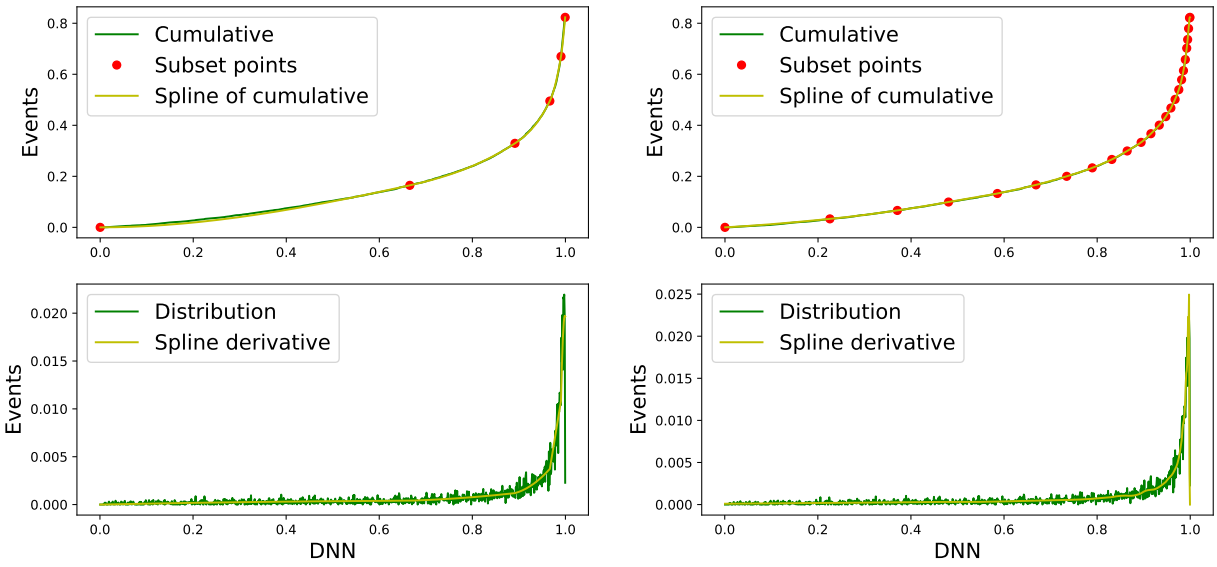
$$\text{error(bin)} \quad = \quad \sqrt{\sum_{\text{mb}\in\text{bin}} \text{var}_{\text{mb}}} \quad . \tag{18}$$

The calculation of those is important for constraints described in Section 6.4.3.

6.4.2 Spline limit optimization

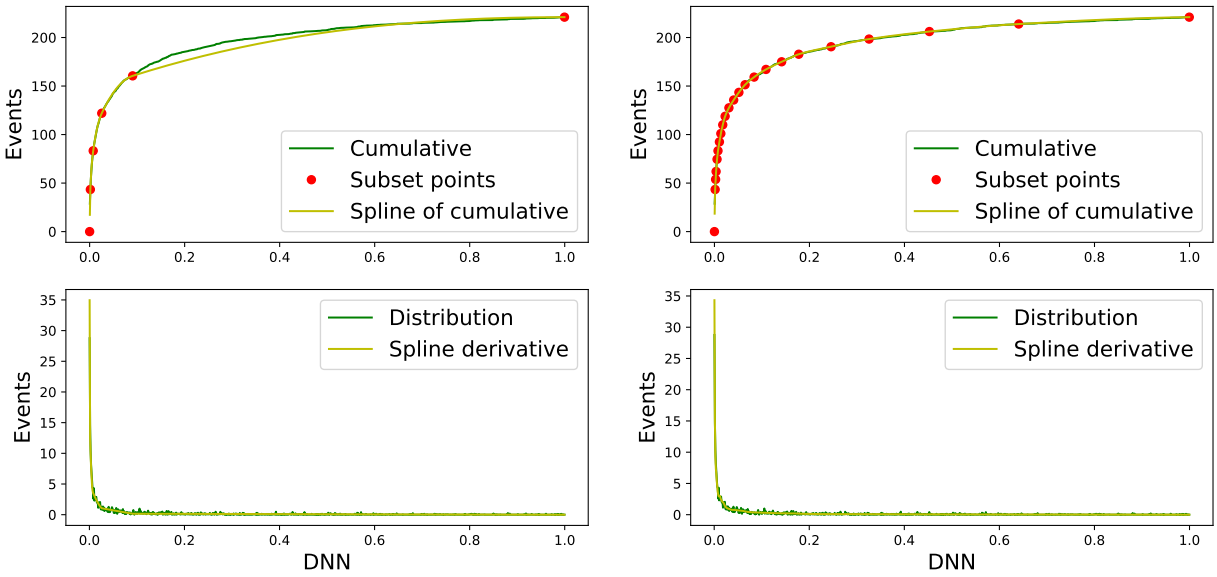
This method is based on the idea not to use the data and mini-bins directly, but to create an analytic representation of those for the optimization. Advantages are, that unwanted statistical fluctuations in the mini-bin contents can be suppressed when fitting the mini-bins, and if the fit is an analytical object, standard optimization tools can be used instead of the brute force calculations in Section 6.4.1. Also the positions of bins are not limited to the discrete width of mini-bins anymore, but can be moved continuously.

Because the fit smooths over the distribution, a trade-off between capturing the overall shape of the distribution and capturing the fluctuations arises. The fit should be precise enough to be a valid representation of the underlying distribution, but should be limited in the number of fit parameters in order not to capture the statistical fluctuations. To achieve that, instead of fitting the distribution of mini-bins, the cumulative distribution over the mini-bins is used. If the underlying distribution has no negative entries, the cumulative is monotone, simpler to fit and allows to use the following method.
Running a fit over the cumulative and using all points of the distribution as parameters would still contain all the unwanted fluctuations. Thus only a subset of the points is used. To select this subset the cumulative is split in a certain equidistant number of intervals in the direction of the number of entries (y-axis). This number of intervals is further referred to as the touch-rate. The maximum of the cumulative divided by the touch-rate is the length each interval should have in the y-direction. Going from zero upwards in y-direction, each point of the cumulative at the end of an interval is used for the subset. Figure 9 shows example distributions. This subset of points describes the steeper parts of the cumulative more detailed than flatter ones. The subset points are then used to fit the cumulative. The first and the last points of the cumulative are always given extra, if not already included, as fit constraints, which can cause the final set of points used for the fit to differ from the given touch rate. Furthermore processes with high amounts of empty mini-bins can have overall fewer points than demanded by the touch rate.
For the fit Piecewise Cubic Hermite Interpolating Polynomials (PCHIP), which is a so called cSpline, implemented by the scipy package [15], are used. In this thesis they are referred to as splines.
Figure 9 shows construction examples of the splines and how they compare to the original distributions. The figure shows in the upper part of each subfigure the cumulative with its subset points the spline uses. The lower plot in the subfigures show the original distribution and the derivative of the spline from the cumulative in the plot above. It shows directly how the mini-bin distribution is represented by it. The derivative is scaled to the same integral as the distribution. The gaps in the upper plots of Subfigures 9(c) and 9(d) are plotting artifacts and actually reach down to the subset point at (0,0).
This method produces a representation of the mini-bin distribution with fewer fluctuations. The scaled derivatives of the splines compared to the original distributions of the original data for all processes and touch-rates of five and 25 are displayed in Figures 37 and 41 in the appendix.



(a) Spline construction for a signal process with touch-rate 5.

(b) Spline construction for a signal process with touch-rate 25.



(c) Spline construction for a background process with touch-rate 5.

(d) Spline construction for a background process with touch-rate 25.

Figure 9: Spline construction examples and comparison to original distributions of one signal process and a significant background process for different touch-rates.

Yields of bins can be recovered by subtracting the splines for the yields (yspl) at the upper bin edge position (e_i) from the fit at the lower one

$$\text{yield}\,(\text{bin}) = \text{yspl}\,(e_i) - \text{yspl}\,(e_{i-1}) \quad .$$

The corresponding errors cannot be calculated directly anymore, because the bin positions are continuous. Therefore the errors are also represented by splines. The error splines are produced by taking the subset points used for the splines and calculate the respective cumulative variances. The error splines (varspl) then fit the x-positions and the cumulative variances. Calculating errors is then done by

$$\text{error}\,(\text{bin}) = \sqrt{\text{varspl}\,(e_i) - \text{varspl}\,(e_{i-1})} \quad .$$

Using the splines instead of mini-bins the calculation of ΔL can be transformed such that it is not dependent on $\Delta s, \Delta b$ but the bin edge position e_i. Here the transformation is shown for any process (p) in the i-th bin with the upper bin edge position e_i and a possible new upper bin edge position e_i'

$$\Delta p_i\,(e_i, e_i') \quad = \quad \mathrm{spl}\,(e_i') - p_{i-1} \quad .$$

With this, the change of limit for a bin and process is only a function of the new bin edge position given the old position, as well as the yield of the old bin and the gradients

$$\Delta L \quad = \quad \Delta L\left(e_i' |\ e_i, p_i, \frac{\partial L}{\partial p_i}\right) \quad . \tag{19}$$

This function is optimizable with standard tools like the minimizer from the scipy package.
The underlying optimization method SLSQP is used, because this method is able to include boundaries and constraints.

The algorithm from Section 6.4.1 is used also for this spline algorithm again with corresponding changes to the calculation in the second step, and the dependencies in the third.

6.4.3 Stepping algorithm

The stepping algorithm is the last step in a round of optimization. It executes how the calculated optimal bin moves are applied.
Because the optimal moves of each bin edge are calculated with all other edges frozen, applying all calculated moves would differ from the predictions. The gradients on the bin yields are also only valid locally, thus moving a bin too much should be constrained.

The locality of derivative is also the reason to recompute the limit after each round of optimization. Therefore the stepping algorithm takes the locality, as well as all the other constraints into consideration.

Apart from the locality of the gradients, each binning should be physical and fulfill statistical constraints to ensure that the statistical inference is stable and robust. Thus the interchange of any two bin edges is prohibited and the maximum step-size itself is limited in the calculation of the limits. In this thesis the maximal step-size for the mini-bin method was arbitrarily set to five percent of the DNN output range. The spline method is restricted by arbitrarily setting the boundaries for edge movements to half of the distance towards the neighbouring edges.

The Constraints

For the statistical constraints at least one MC event has to be present in each bin of the two most prominent background processes (ttbar and Drell-Yan (DY)). Also the relative MC error of all

background processes combined in a bin should be below 70%. For the relative MC error the QCD contribution is excluded, as its error is not MC driven.
Without these constraints it is optimal for the limit to have large signal to noise ratios, leading the algorithm to aggressively remove background in the signal heavy bins until only signal is left. This corresponds to the rightmost bin of the classifier to be extremely narrow.

The Algorithm

The stepping algorithm consists of the following steps for one round of optimization iteration:

1. Apply the bin edge move with highest impact on the limit.
2. Freeze bin edges adjacent to bins who experienced change due to a previous edge move.
3. Freeze bins that would move already frozen bins.
4. If possible moves remain, start from 1. step again. Otherwise, take the new binning for a new limit calculation.

An example schematic can be seen in Figure 10.
This algorithm performs always the most relevant movement and allows for multiple moves per limit calculation. It breaks down the N dimensional problem into N separate 1-D problems, which allows for this simple optimization scheme. In comparison to a multi-dimensional optimizer, the optimization uses less of the calculated gradients and would therefore perform less efficient per round. However, the important part is the need for recomputing of the limit after each round. It improves the limit, but increases computing time.

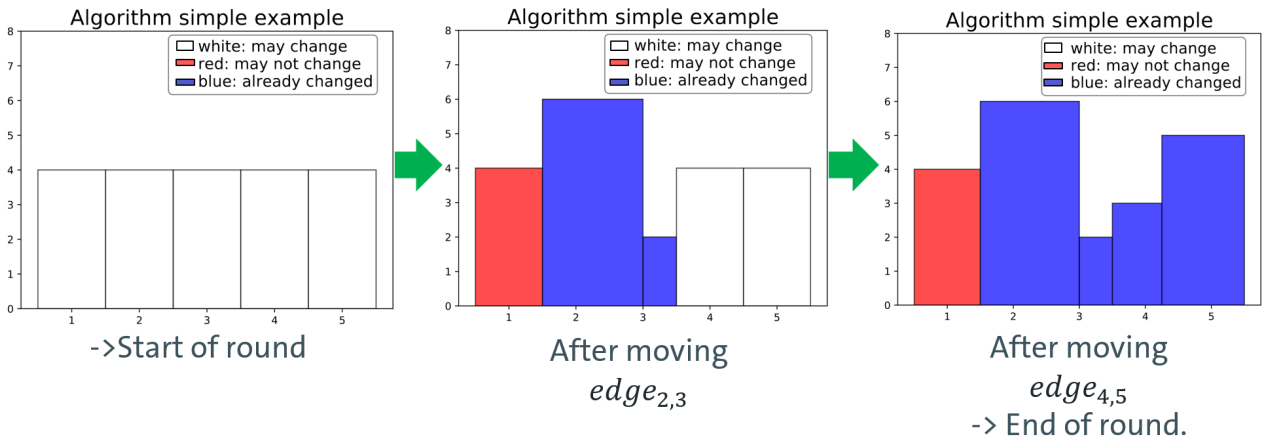


Figure 10: Stepping algorithm to apply possible moves for optimization.

7 Optimization Results

This chapter contains results from the examined methods. First a series of validation tests are discussed for the general methodology which is followed by the results of the two methods.

7.1 Validation Results

As a comparison between combine and pyhf upper limit scans of statistical models with the same uncertainties and fixed binning are performed. The results can be seen in Figure 11 for pyhf and Table 1 for combine.
Differences are expected due to different methods in model building and and especially the handling of the MC errors, but in the shown scans the MC errors are not included. As mentioned in 6.3, the limit calculations in this thesis are not using the statistic errors. Thus the expected limits are very similar. For example the expected 95% upper limit at 50% for the combine scan lies between a poi of 12 and 13, but closer to 12. The pyhf scan interpolates between the calculated μ and yields approximately 12.2974 at 50%, which are denoted as 0σ in the figure. Therefore the pyhf scan is reliable and comparable to combine.

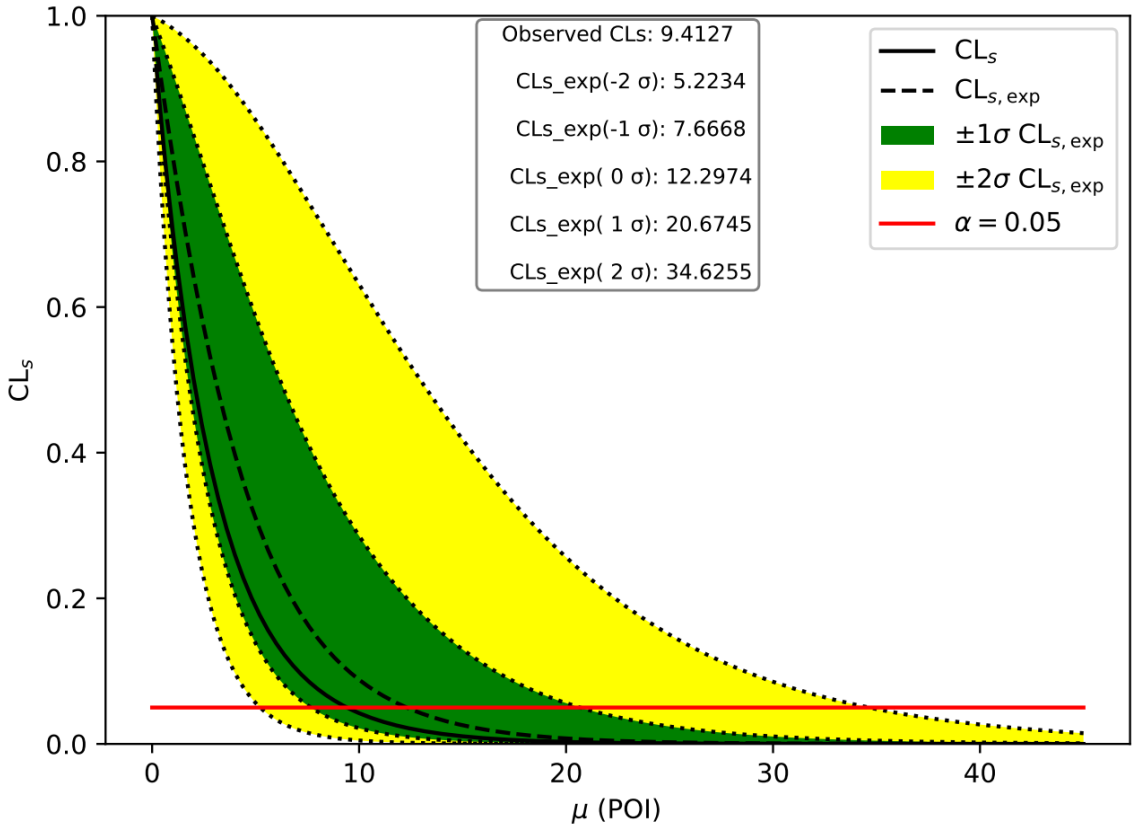
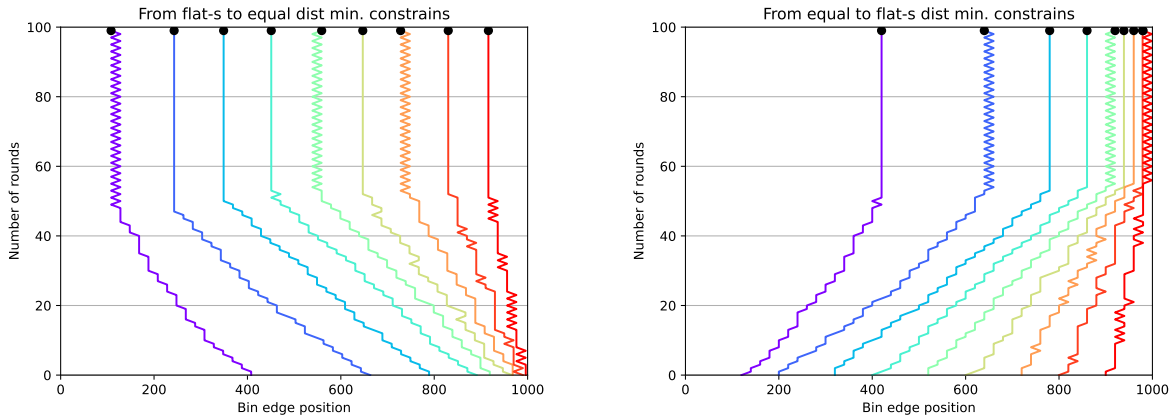


Figure 11: pyhf CL_s scan. The model does not include the statistic errors.

Expected	μ	CL_s
2.5% :	5	0.06059
2.5% :	6	0.03650
16% :	7	0.068
16% :	8	0.04729
50% :	12	0.05582
50% :	13	0.04351
84% :	20	0.05767
84% :	21	0.04821
97.5% :	34	0.05143
97.5% :	35	0.04574

Table 1: Table with values from a combine CL_s scan. The scan uses the same model configurations as the pyhf scan in 11.

As a closure test for the underlying methodology, test gradients are used with the mini-bin method. These gradients are set up to lead towards a dedicated target binning. For this test an equal binning and a flat-s binning are used as starting and target binning and tested in both directions. Flat-s is defined as having the same amount of signal events in each bin. The result is shown in Figure 12. In the figure the progress of the binning over multiple rounds of the optimization is plotted. The lower x-axis indicates the bin edge positions in the DNN output space in terms of mini-bins. The colored lines are the bin edges, with the black dots signaling their final positions. The y-axis, starting at the bottom, indicates the number of rounds of optimization done.
The gradients are the difference to the target positions in mini-bin terms. The binning fluctuates around the final positions, because the method does not lower the maximum step-size over time, but uses convergence conditions e.g. a binning repeating multiple times. Furthermore, minimal constraints and boundaries are enforced: each bin has to have at least one mini-bin and edges can not cross.



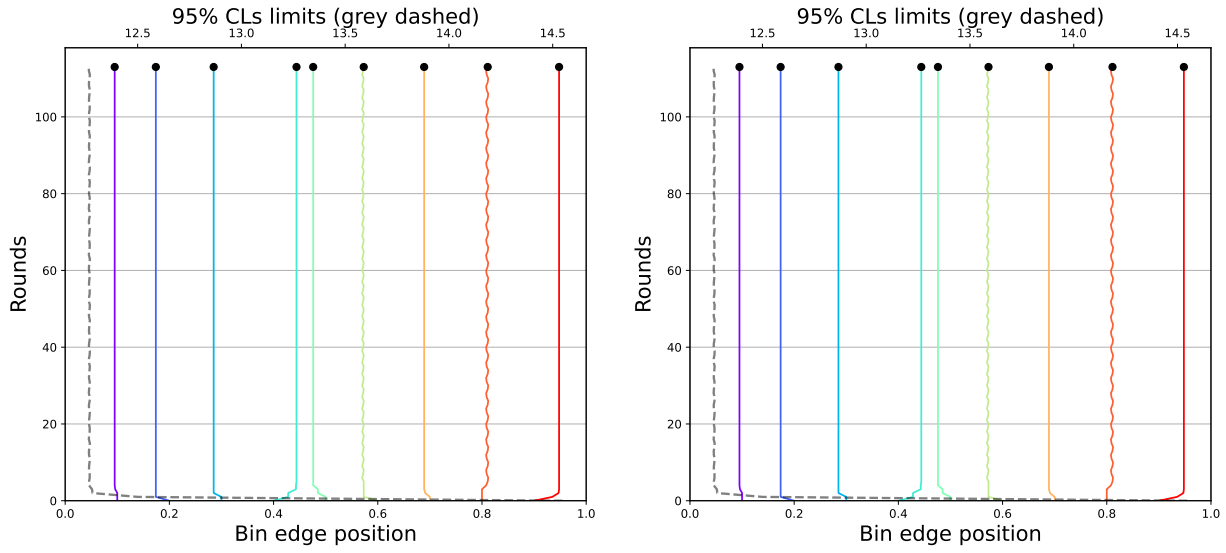
(a) Test gradients to guide a flat-s binning to an equal distance binning.

(b) Test gradients to guide an equal distance binning to a flat-s binning.

Figure 12: Bin edge position history for test gradients to achieve a certain binning. The bin edge position is given in the amount of mini-bins on the x-axis. The rounds of optimization are given on the y-axis.

7.2 Mini-bin Results

To evaluate the mini-bin method, 500 toys are produced as variations of the original data. For that, each new mini-bin yield of each process is drawn from a log-normal distribution with the original value as mean value of the log-normal and the statistical MC error of the original value as standard deviation of the log-normal. The log-normal distribution is used to not draw negative yields, which would be likely for empty mini-bins and a normal distribution, as this would be non-physical.
Then the optimization scheme from Section 6 is used to optimize the binnings of the original and all toy distributions. This is done once with and once without enforcing the statistical constraints. The optimization always starts from an equal-distance binning, because it is a simple binning that also fulfills the constraints. The number of bins is arbitrarily set to ten.
In Figure 13 history plots of the optimizations are shown for original data with and without constraints respectively. The bin-edge position in the DNN output space is indicated on the lower x-axis for the colored bin edges. The number of rounds of optimization performed are on the y-axis. The gray dashed line indicates the limit for each binning in each round with its values on the upper x-axis.
Both optimizations perform the same. After a few moves in less than ten rounds a binning is reached where only minor changes are applied for the rest of the optimization. Convergence is defined as repetition of a configuration ten times.

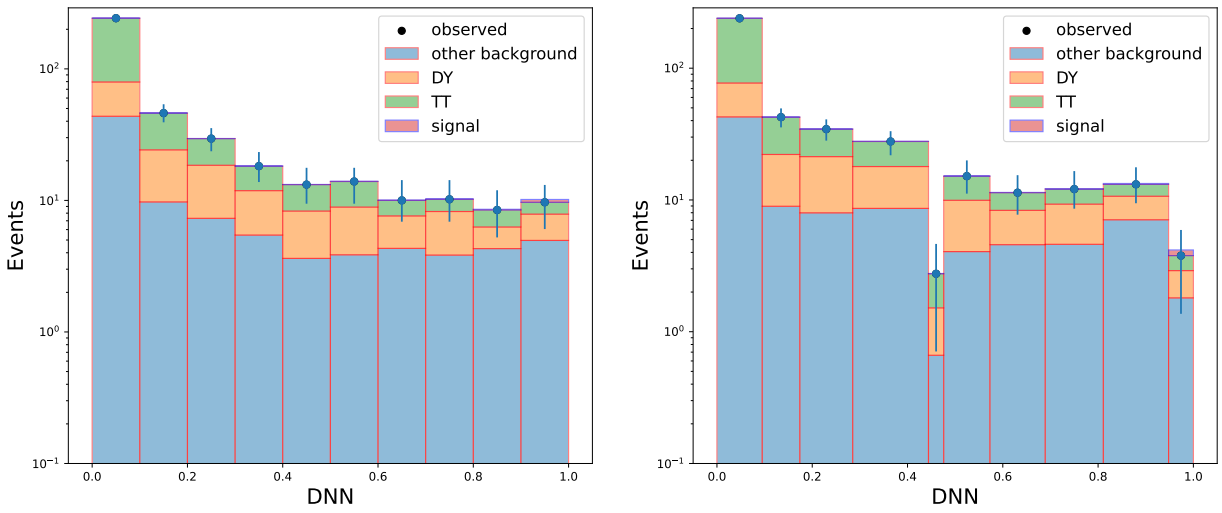


(a) No constraints enforced

(b) Constraints enforced

Figure 13: Optimization binning history with and without statistical constraints enforced on the original data. The grey dashed line refers to the 95% CL_s limit on the upper x-axis. The colored lines represent the bin edges on the lower x-axis.

Figure 14 shows the histogram with the start binning and the optimized binning.

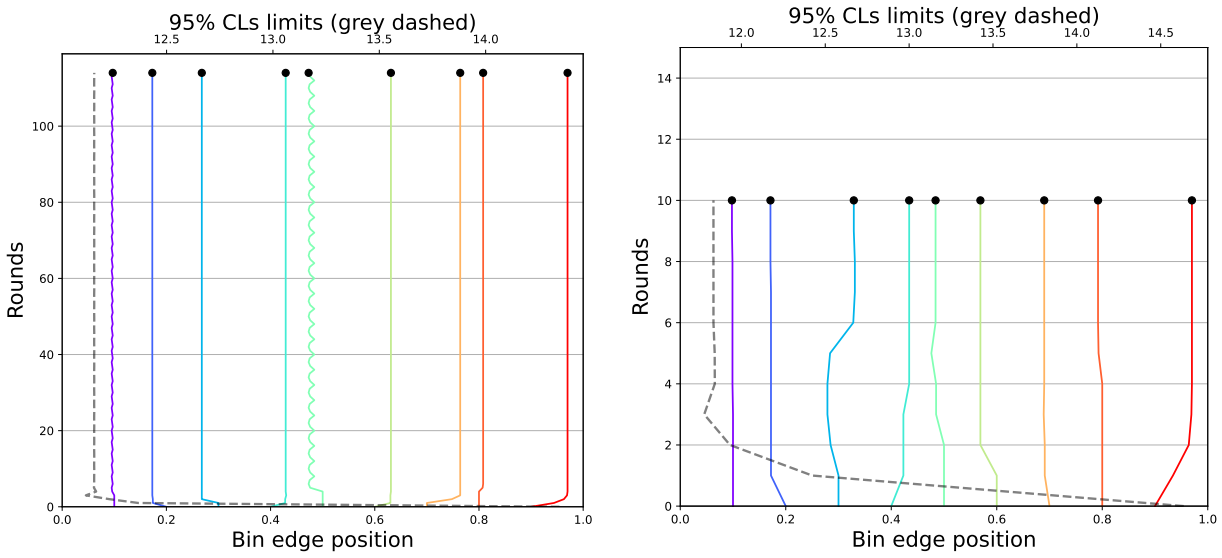


(a) Histogram of data with equal distance binning.

(b) Histogram of data with optimized binning.

Figure 14: Original data binned with equal distance 14(a) and with the constrained optimized positions in 14(b). The observed data is the sum of all background events.

The mini-bin method is not protected against "bad" steps where the limit increases. In Figure 15 two optimizations of different toys are shown where such steps happen after a few rounds and the method is not able to return to these positions, that have a smaller limit then the final positions, before converging.



(a) Toy number 1

(b) Toy number 5

Figure 15: Optimization binning histories without statistical constraints enforced on the two toys. The grey dashed line refers to the 95% CL_s limit on the upper x-axis. The colored lines represent the bin edges on the lower x-axis.

To evaluate how the variations in the data affected the optimization, Figure 16 depicts the distributions of limits for data and all toys. It contains the limits for the equal start binning (in blue) together with the final limits after constrained (in red) and unconstrained optimizations (in orange). The limits of the original data are marked with black vertical lines. The skew in the distributions is likely to arise from using the log-normal distributions for the toys and it seems to become less after

the optimizations. The constraints seem to impact the optimizations only to a small extent, as the distributions are very similar. The standard deviations imply the average impact the variations of the toys have on the limit. This is reduced after the optimization and is a little bit smaller for the distribution with constrained optimizations.

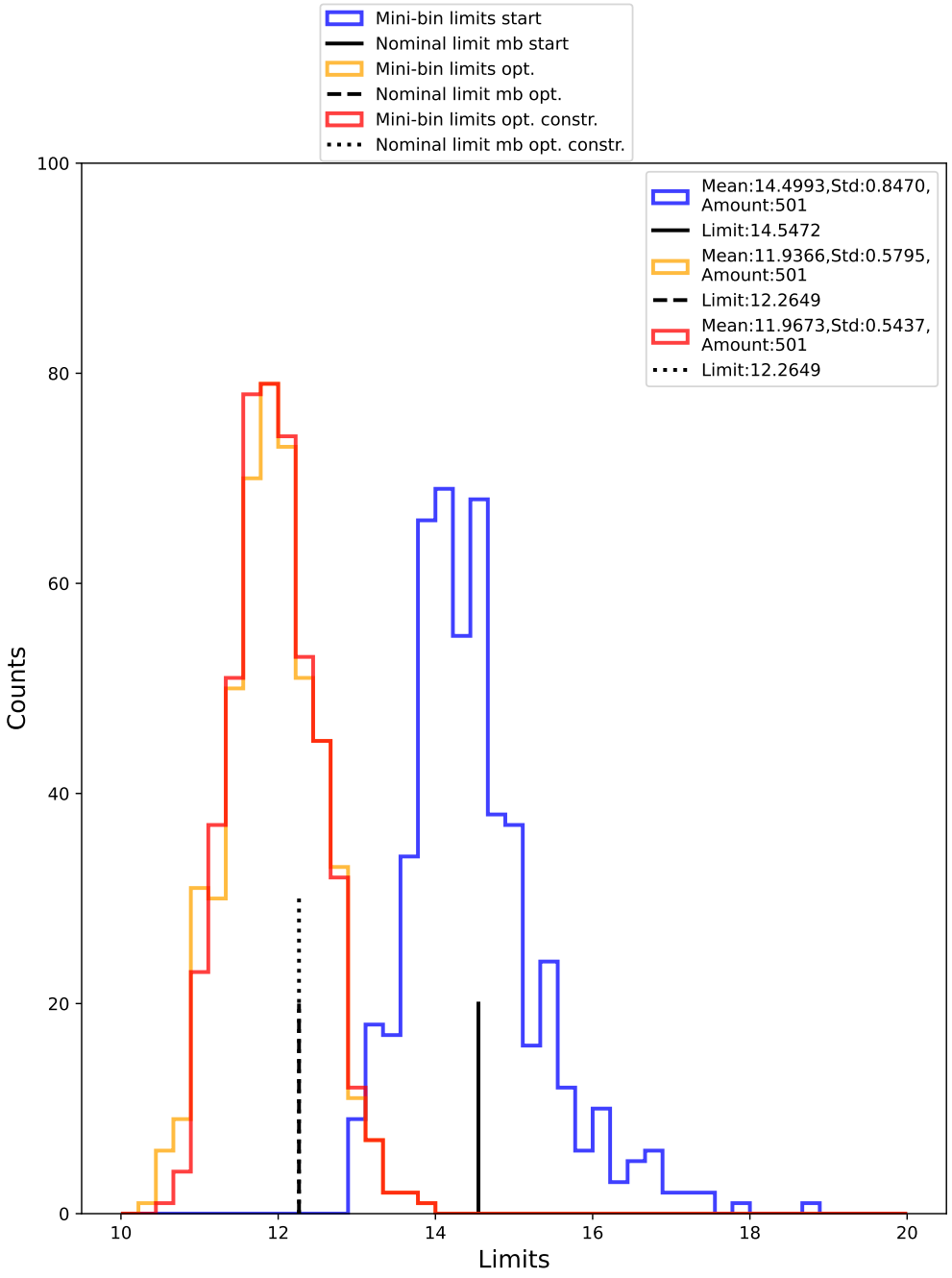


Figure 16: Distributions of limits for data and toys. The mean values (mean), standard deviations (std) and the amount of data points (500 toys plus original) are shown in the legend. The starting binning is an equidistant binning.

Figure 17 shows for each toy and the original data the difference between the limits before and after optimization. The average improvement is about 2.5 for the toys and similar for the data from which the toys are derived. This improvement is significant in comparison to the average value of 14.5 before optimization. The width of the distributions relates to the influence of the variations in the toys on the optimization. A small width would imply that the toys limits are optimized approximately the same amount. The widths are relative to the average improvement

approximately 30% for the constrained and unconstrained optimization, which is significant.

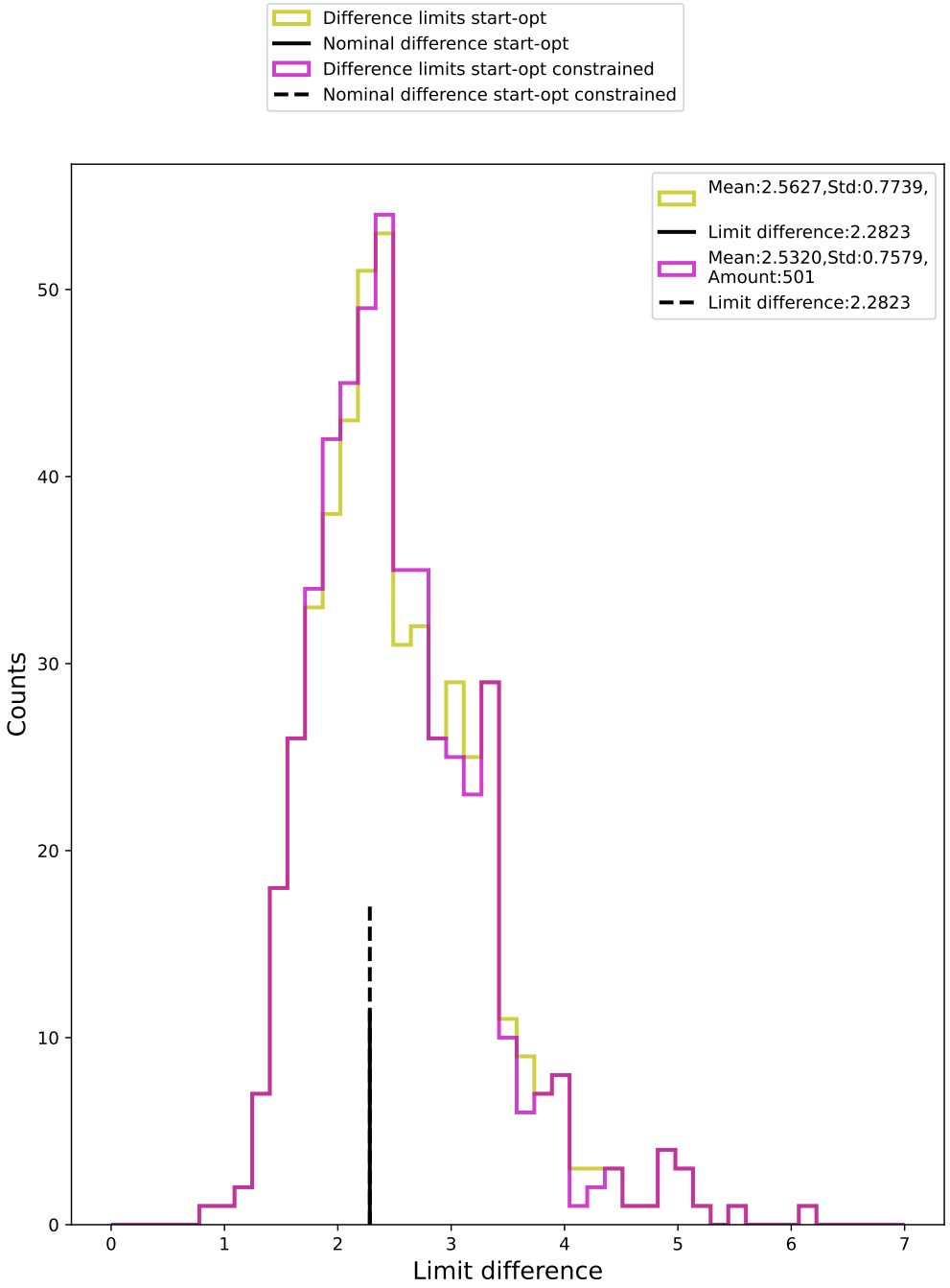
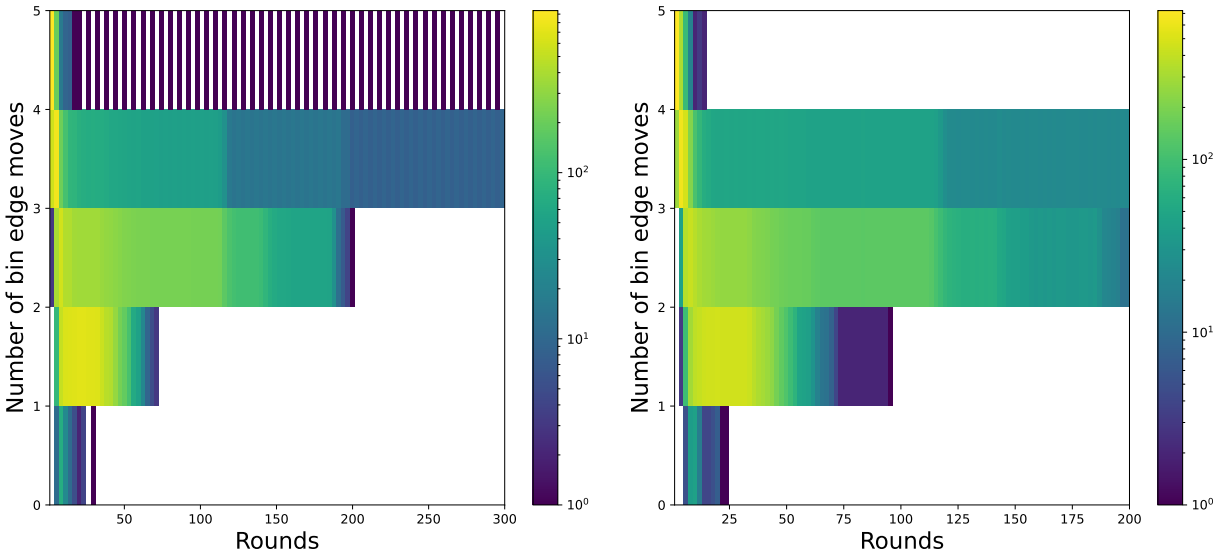


Figure 17: Distributions of differences between limit before (equidistant binning) and after optimization, with (magenta) and without (yellow) constraints. The original data is highlighted in black. The mean values (mean), standard deviations (std) and the amount of data points (500 toys plus original) are shown in the legend.

Finally Figure 18 depicts the amount of bin-edge movements in certain rounds of the optimizations. Up to four moves can be performed in one round with nine bin-edges and the freezing of neighbouring edges from the stepping algorithm as explained in Section 6.4.3. Zero steps are recorded when the gradients vanish and the optimization converges by itself, without extra convergence conditions being triggered, e.g. repeating binning. In the very first rounds three or four edges are moved most often. After that fewer and mostly single moves are done. After approximately 50 rounds a lot of toys converge. Longer optimizing toys have mostly two or three edge movements.

The stepping algorithm is thus able to get close to how multi-dimensional optimizer would perform.



(a) No constraints enforced

(b) Constraints enforced

Figure 18: Amounts of bin-edge moves in certain rounds of optimization.

7.3 Results with Splines

In Figure 19 examples of the splines are shown. A signal process and some background processes are shown for a touch-rate of 5. The whole figure with all processes and figures for other touch-rates are available in the appendix.
The splines fit better with a higher touch rate, as expected. However, for processes with many empty mini-bins, such as VVV, deviations are high for all touch rates, because the points for reconstruction are limited by the number of steps the cumulative does. As this effects only small background processes the effect is negligible.
Not to be neglected are relatively small deviations on large background processes like TT or on the signal processes. For the touch-rate of five, in the region around 0.2, the spline underfits the TT cumulative yield. Because the first and last points of the cumulative is always in the spline, the TT events lost in this region are, in this case, assigned by the splines to higher regions around 0.8. For large background processes like TT or any signal process, these shifts in events can be significant for the signal extraction. Thus choosing a touch-rate has to be done carefully.

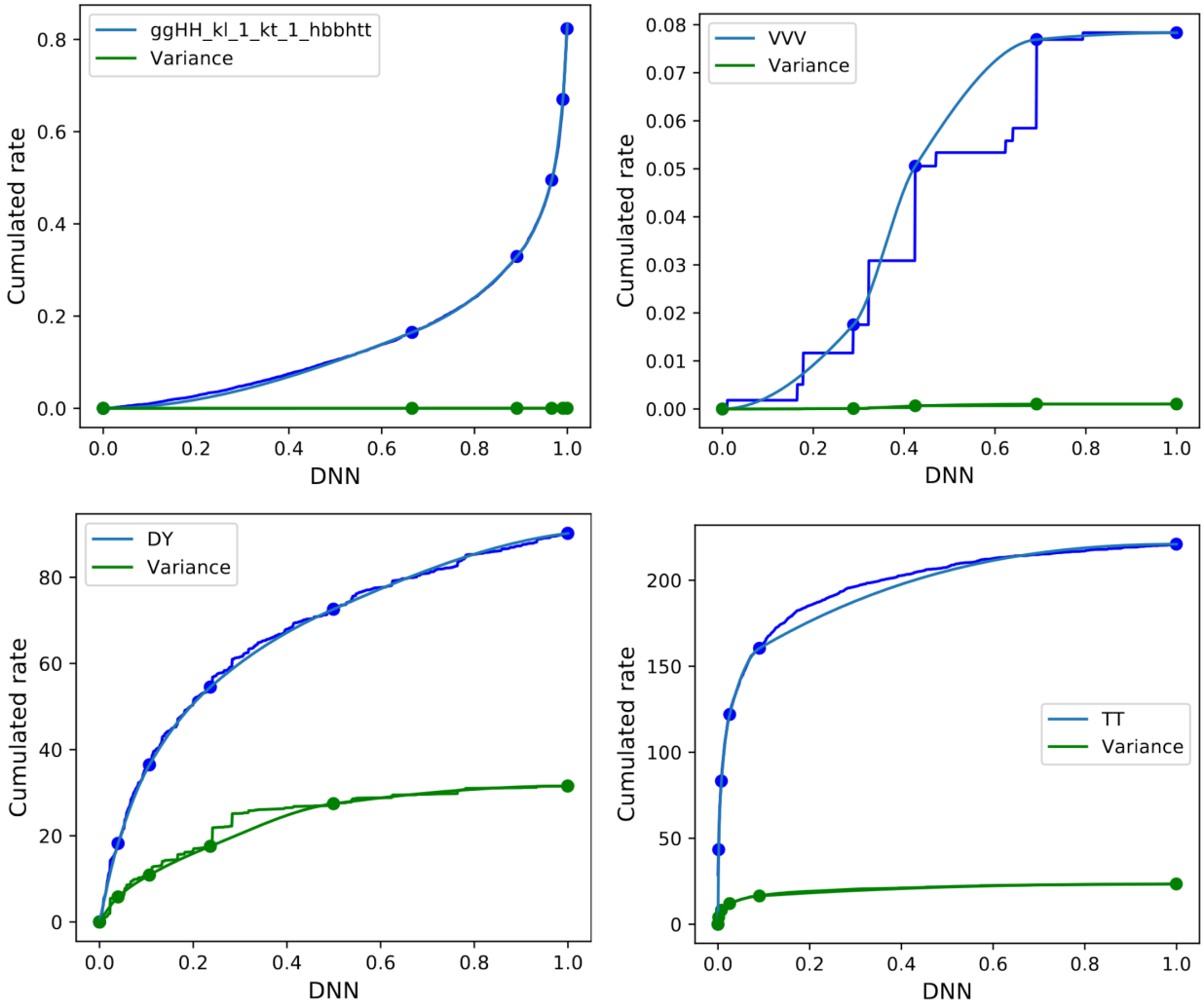


Figure 19: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 5. Shown are for the original data the results for a signal and some background processes (see legends). The complete figure with all processes is available in the appendix (Figure 42).

7.3.1 Splines closure tests

The splines should be stable under variations of the data. In Figure 20 the chi-square per degree of freedom (chi-square/dof or reduced chi-square) distributions of the toys are plotted for a signal and some background processes. The chi-square is calculated by taking the splines as prediction for the yields for a histogram of the classifiers values. In the figure a touch-rate of 10 is shown whereas 10 equidistant bins are chosen for the histogram. The full figure with all processes and figures for the other touch-rates are available in the appendix. For higher touch-rates the distributions show smaller means and standard deviations. The original data value is marked with a black vertical line again. Most mean values of the distributions are on the order of one or below. The reduced chi-square for the original data lies often below the means. Reduced chi-squared values of one or below are considered sensible for the splines. Very small values are not to be considered as over-fitting here, because the distribution is not fitted directly, but already approximated with the subset of points. The standard deviation values further indicate how much influence the variations have on the splines performances.
Very big or extremely small mean or standard deviation values can be seen for processes with small rate, as they are more often fitted either very precise or very inaccurate, e.g. singleT and W.

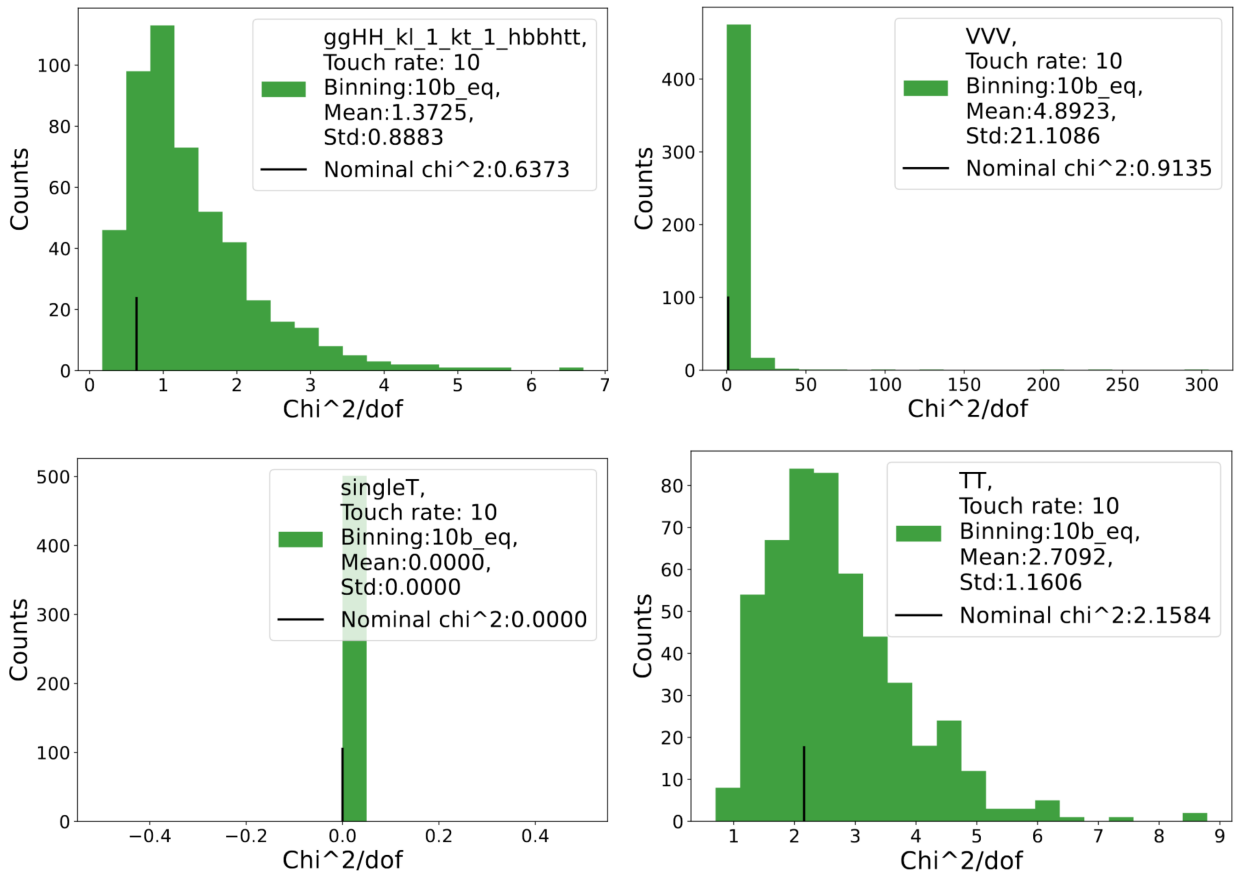
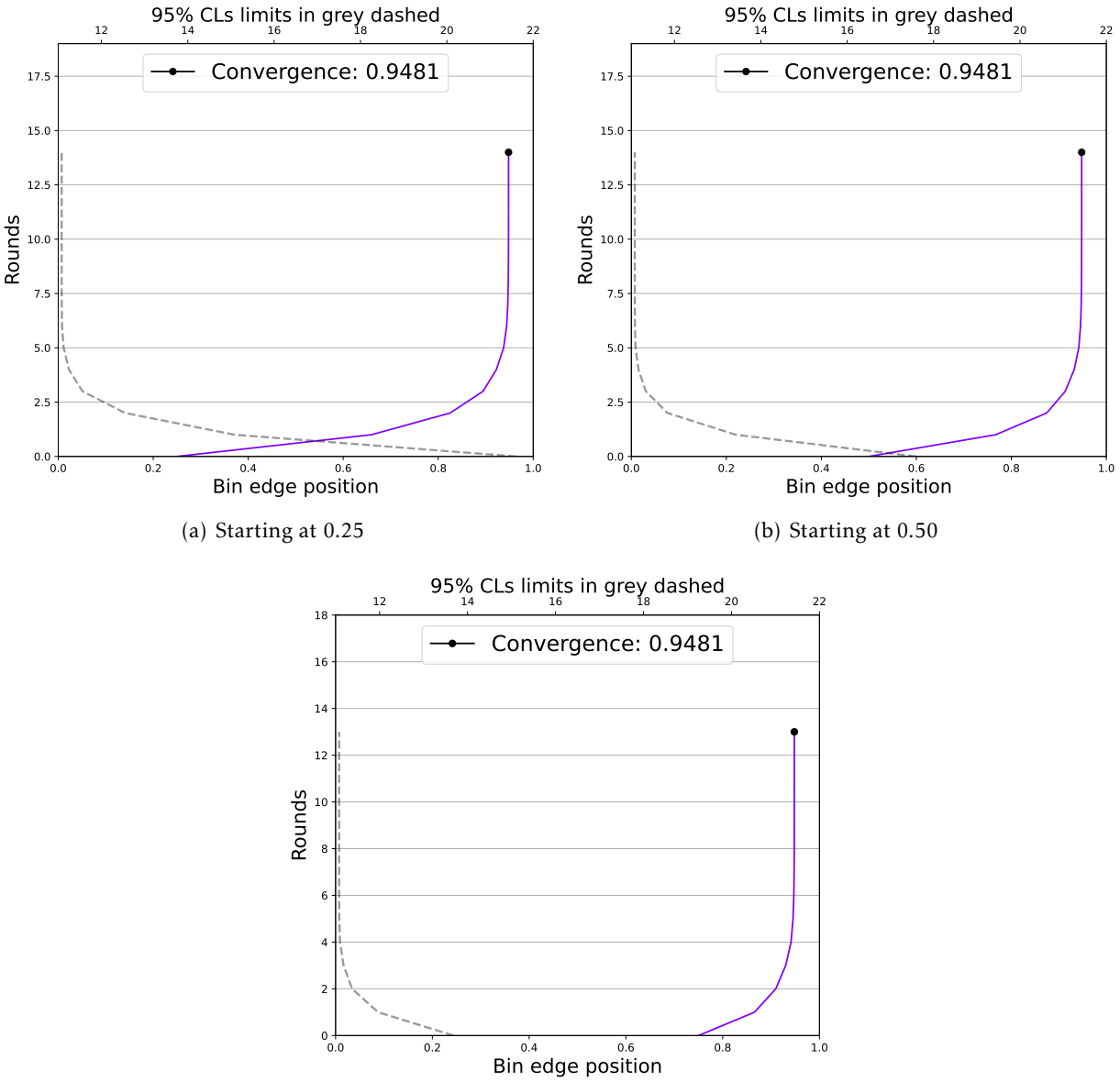


Figure 20: Distributions of reduced chi-squares (Chi^2/dof) for all processes of the splines for a touch-rate of 10 and an equal binning with 10 bins (10b_eq). Shown are toys and original data in green for a signal and some background processes (see legends). The original data is highlighted in black. The full figure with all processes is available in Figure 53 in the appendix.

To test the spline optimization, optimizations on a binning with two bins and therefore one bin edge can be seen in Figure 21. No statistical constraints are enforced. The three different starting positions converge at the exact same position of 0.9481.



(a) Starting at 0.25

(b) Starting at 0.50

(c) Starting at 0.75

Figure 21: Optimization of a single bin edge/ two bins for different starting positions. Upper limits of each binning is plotted as a grey dashed line with the values on the upper x-axis. The purple lines represent the bin edge position in each round on the lower x-axis. The labeled convergence points show the positions the edges converged to.

Finally Figure 22 depicts how the optimization would perform without recomputing the limit after every round. In the first three rounds minor positive changes are visible in binning and limit, before a slight movement in round four pushes the limit to a slightly higher value than at the start, at which it stops moving. More sophisticated optimizers can extract more information from the original limit by building higher derivatives during the optimization, but only for single rounds. Thus the advantage they could have should be minor.
A new limit computation is needed either way, becoming the main computational bottle neck, that cannot be resolved otherwise.

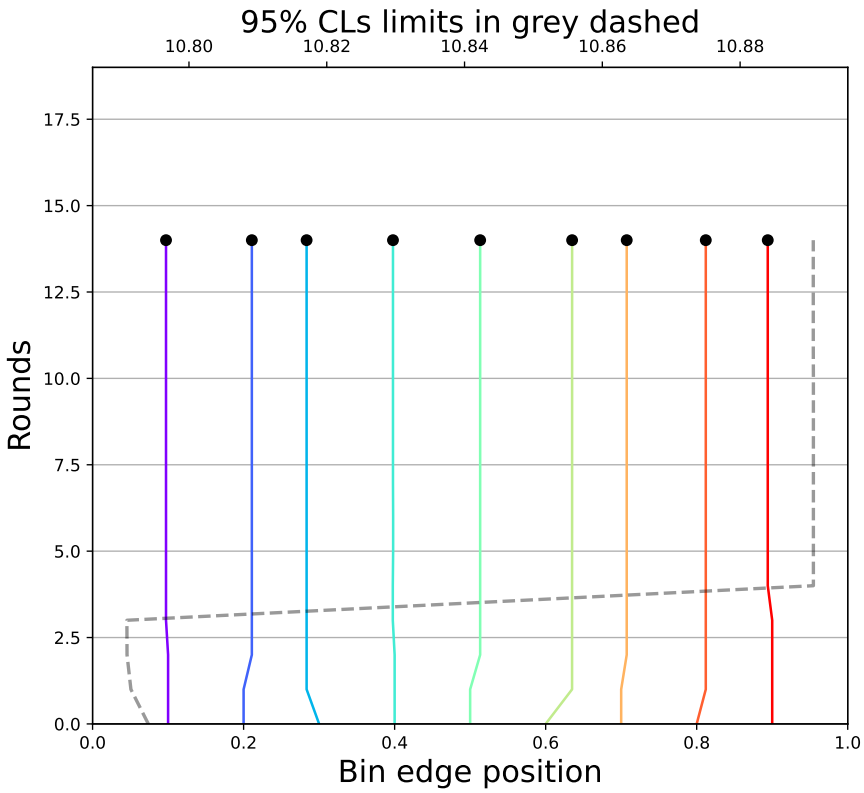
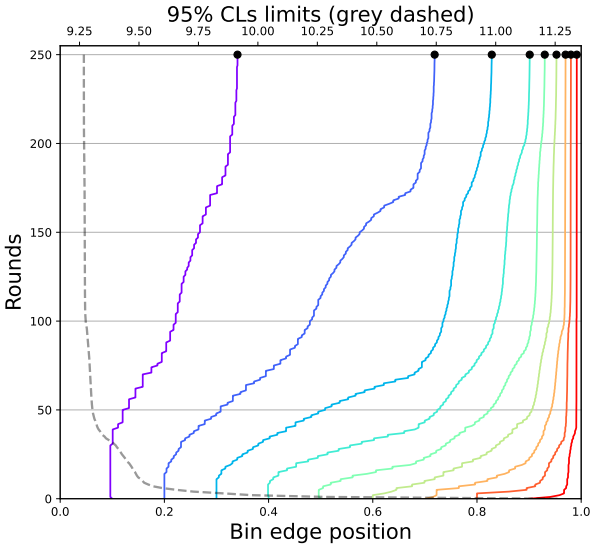


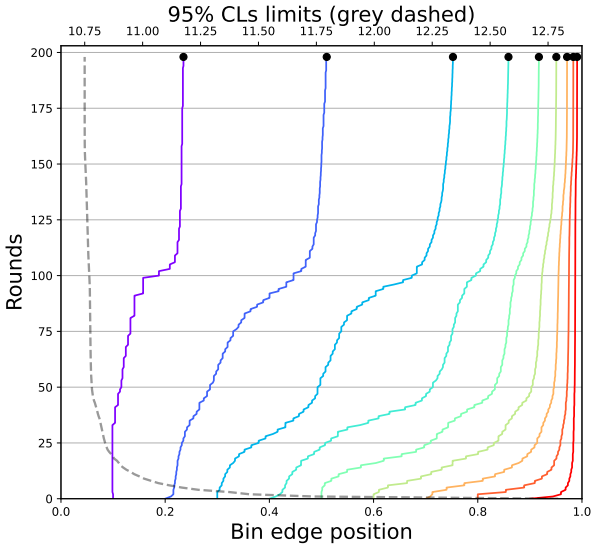
Figure 22: Optimization history without recomputing the limit. The grey dashed line refers to the 95% CL_s limit on the upper x-axis. The colored lines represent the bin edges on the lower x-axis.

7.3.2 Spline optimization without constraints

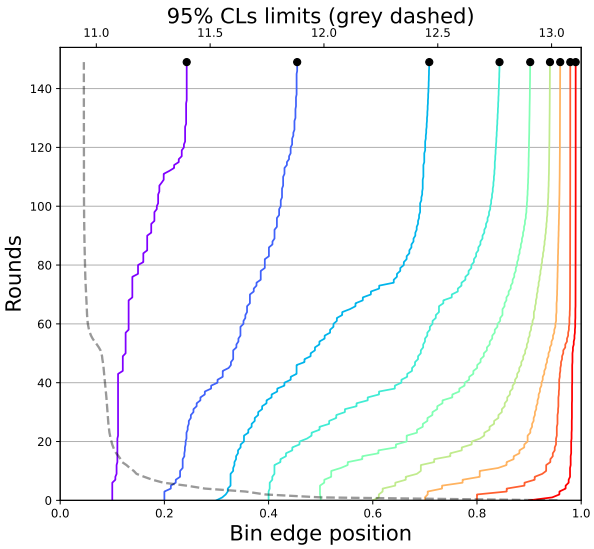
First the method is tested without the statistical constraints being enforced. Figure 23 shows that the method is then able to over-optimize the binnings. The most significant bins at higher values are optimized such that they are as small as possible to reduce the amount of background as much as possible and also the upper limit. The algorithm is not protected against this. This is expected when the constraints are not enforced.
Because the stepping algorithm chooses the moves with largest improvement of significance first, the bin-edge furthest to the right, is optimized first and then the second most significant edge and so on.



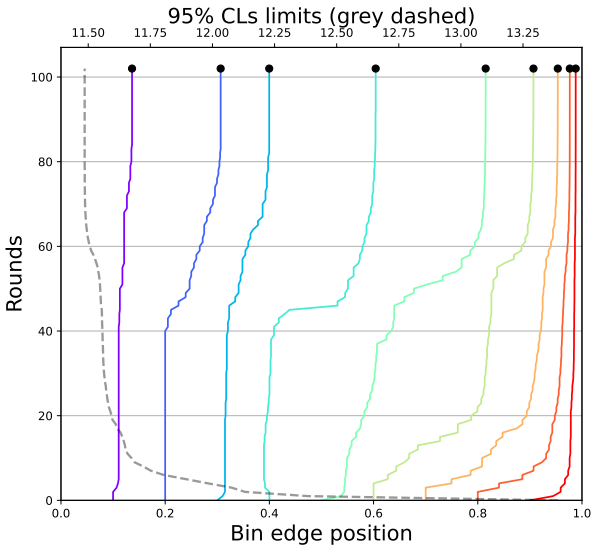
(a) Touch-rate 5



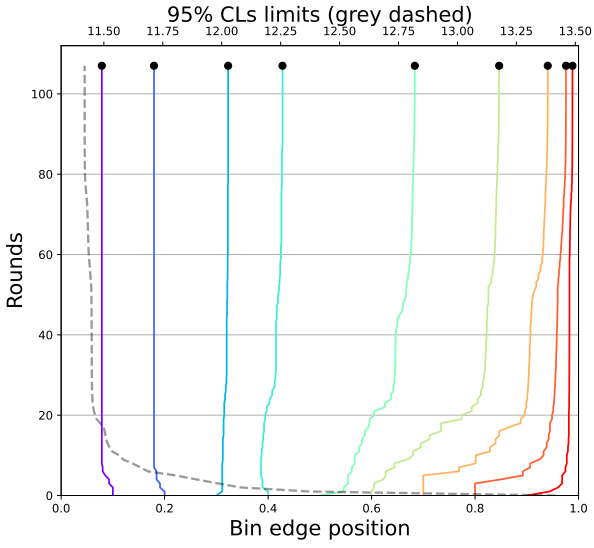
(b) Touch-rate 10



(c) Touch-rate 15



(d) Touch-rate 20



(e) Touch-rate 25

Figure 23: Optimization of the original data for different touch-rates without constraints enforced. The grey dashed line refers to the 95% CL_s limit on the upper x-axis. The colored lines represent the bin edges on the lower x-axis.

In Figure 24 the distributions of optimized limits based on splines of all the toys are shown for different touch-rates together with the starting limits. The original data is indicated with black vertical lines.

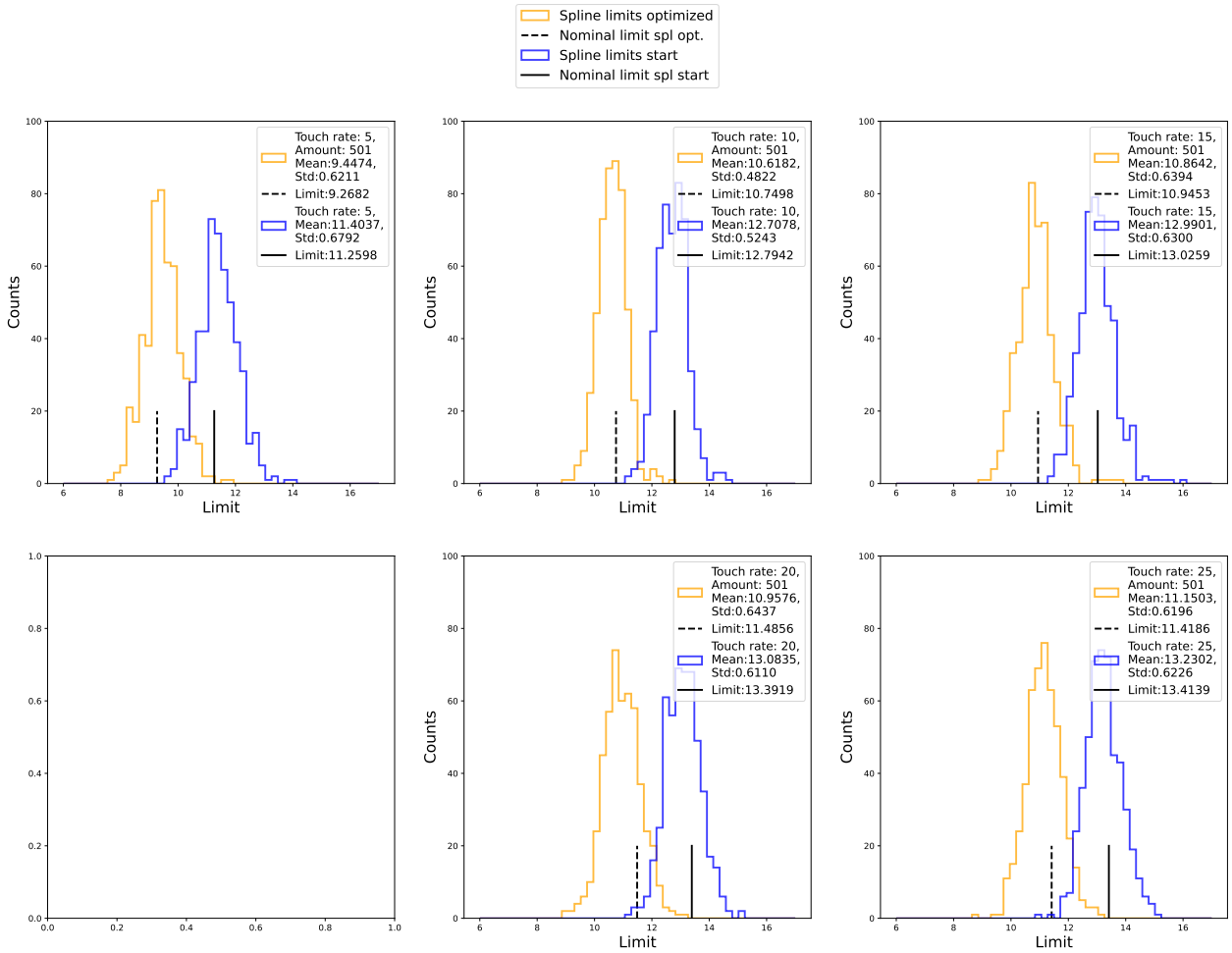


Figure 24: Spline limit distributions with and without optimized binning without constraints for different touch-rates. The mean values (mean), standard deviations (std) and the amount of data points (500 toys plus original) are shown in the legends.

Corresponding to the distributions of limits is Figure 25. It plots the mean limit of the limit distributions from Figure 24 as a function of the touch-rate.
The mean limits before and after optimization are lower for smaller touch-rates due to a worse representation of the splines for the background in the significant bins, leading to less background. For higher touch-rates this seems to be less influential. The mean limits of starting and optimized binnings seem to stabilize for the higher limits, but not definitively. The mean limits are lowered approximately by two for each touch-rate. However, that reduction can not be used for interpretation, as the binnings are not statistically meaningful without fulfilling the constraints.

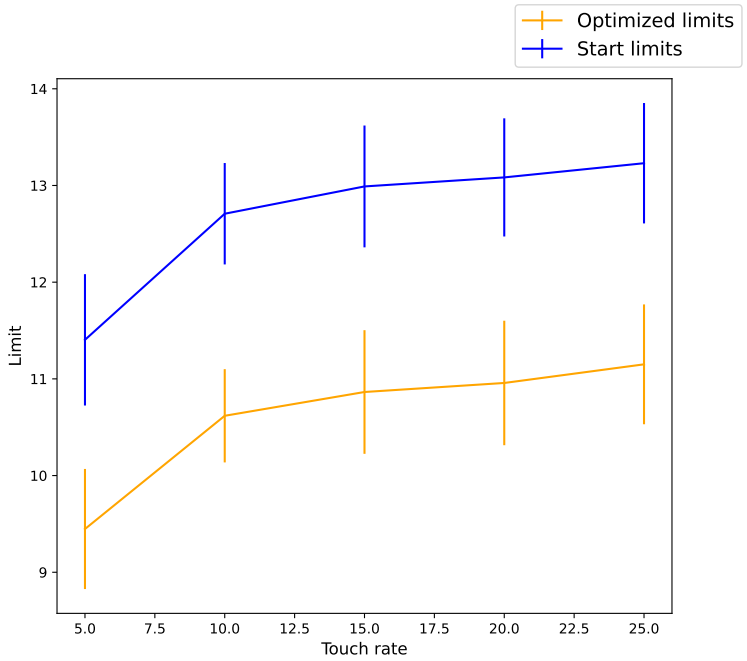


Figure 25: Mean limit as a function of touch-rate for equal distant binning (in blue) and for optimized binning (in orange). No constraints in the optimizations. The error bars denote the width of the limit distributions of original data and the toy experiments.

Figure 26 shows the difference between the limits before and after the optimizations for different touch-rates. The average improvement is for all touch-rates approximately 2. This are significant improvements compared to the average values of 11.4 for the touch-rate of 5 and up to 13.2 for the touch-rate of 25 before optimization. The widths are relative to the average improvements approximately 6% for the touch-rate of 5, 7% for the touch-rate of 10, 9% for the touch-rate of 15, 13% for the touch-rate of 20 and 18% for the touch-rate of 25.

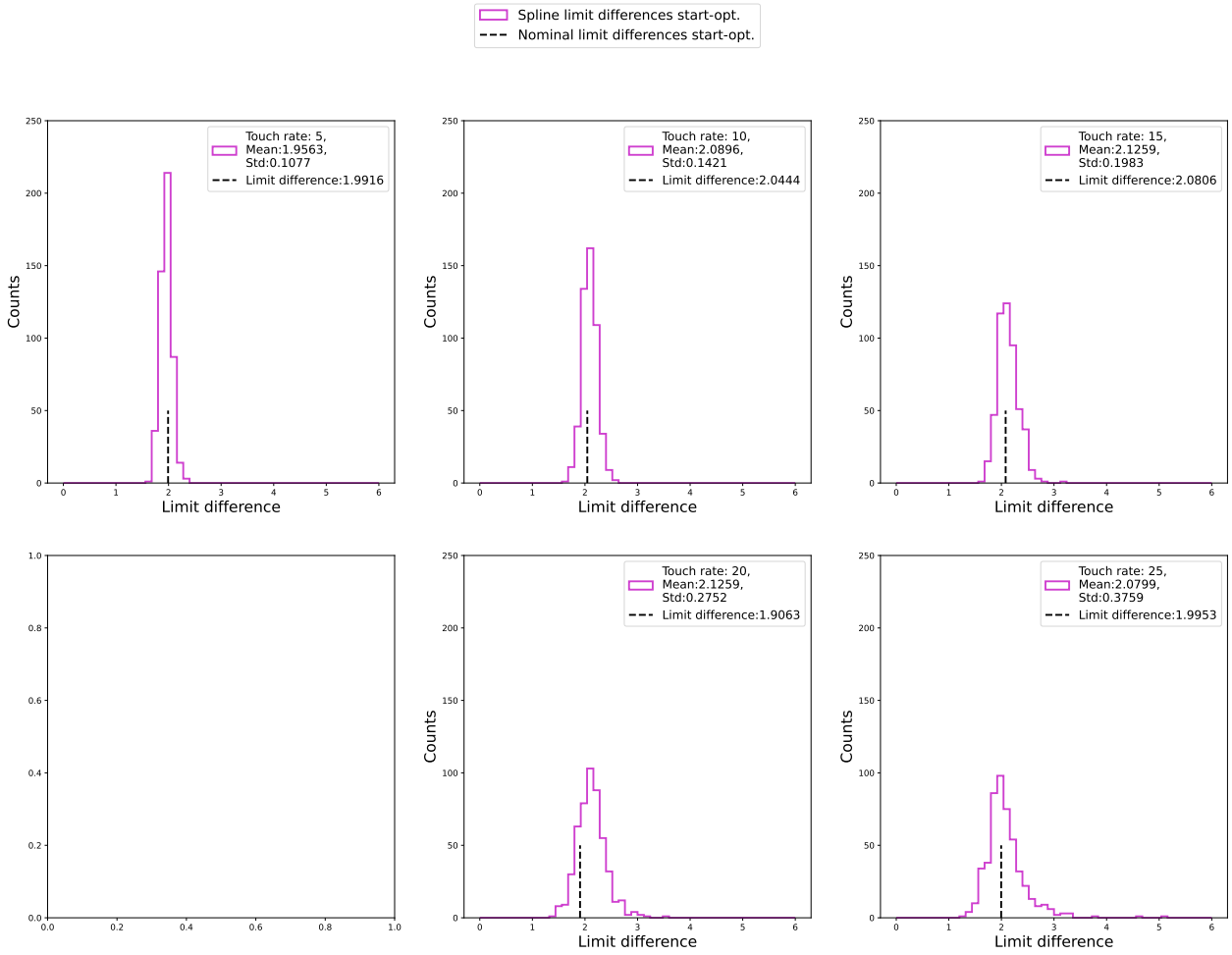


Figure 26: Distributions of differences between limit before (equidistant binning) and after optimization without constraints enforced for different touch-rates. The black line indicates the original data. The mean values (mean) and standard deviations (std) are shown in the legends.

Because the spline binning is continuous, its optimized binning can not be directly compared to values, where the event yields are obtained from the corresponding mini-bin yields. To make a comparison a transformation to the closest mini-bin position is done.
To achieve this, the edge positions are shifted to the nearest of the 1000 mini-bin positions, thus moving each edge a maximal amount of 0.0005. On the resulting transformed binning, the limit is calculated again with the corresponding mini-bin yields.
This is done for the original data, all toys and for each touch-rate. In Figure 27 the optimized spline limits are plotted against its mini-bin transformed limits (Limits mb transform) in a 2-D histogram, with their respective mean and std values. The red line shows the identity transformation.
It approaches the identity for higher touch-rates, thus the spline representation of the bin yields is again closer to the actual yields for higher touch rates.
However, also more limits stay below the identity line for higher touch-rates, which means that the over-estimation of the limit becomes more prominent.

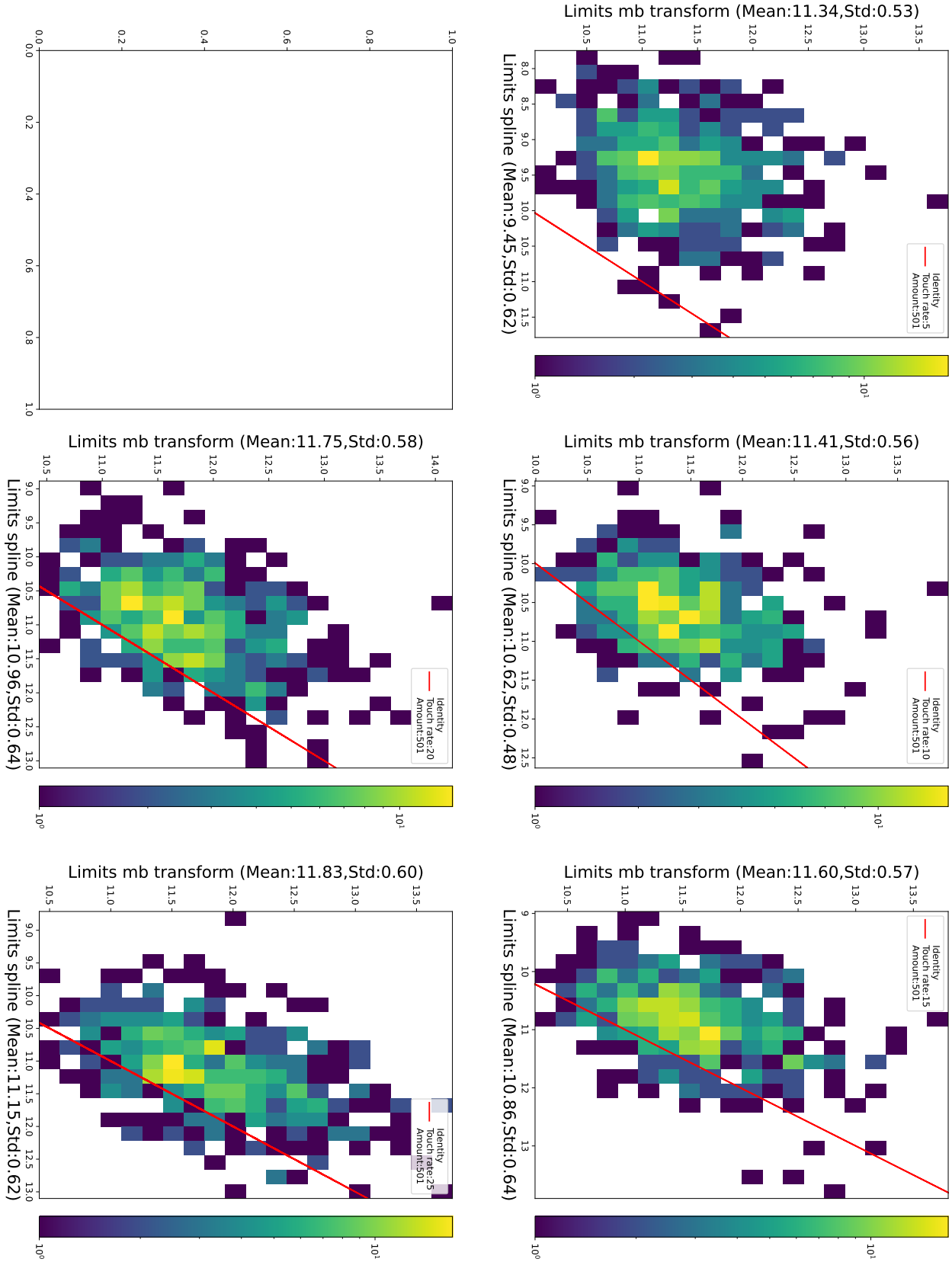
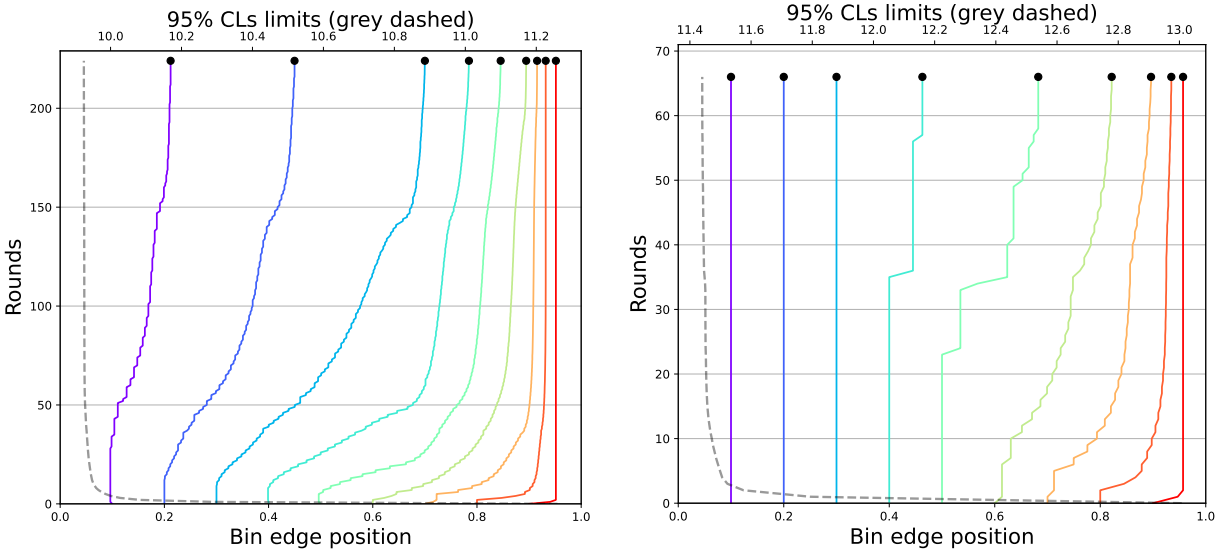


Figure 27: The transformation of optimized spline limits to limits based on mini-bin yields is shown as a 2-D histogram. The red line shows the identity transformation. No constraints are enforced on the optimized splines. The mean values (mean) and standard deviations (std) are shown at the axes labels.

Compared to the mean of the unconstrained mini-bin optimization limit distribution from Figure 16 11.94 ± 0.58, the mean limit of the transformed spline limits is for the lowest touch-rate smaller (-0.60) and comes closer to the mini-bin limits with the highest touch-rate (-0.11).

7.3.3 Spline optimization with constraints

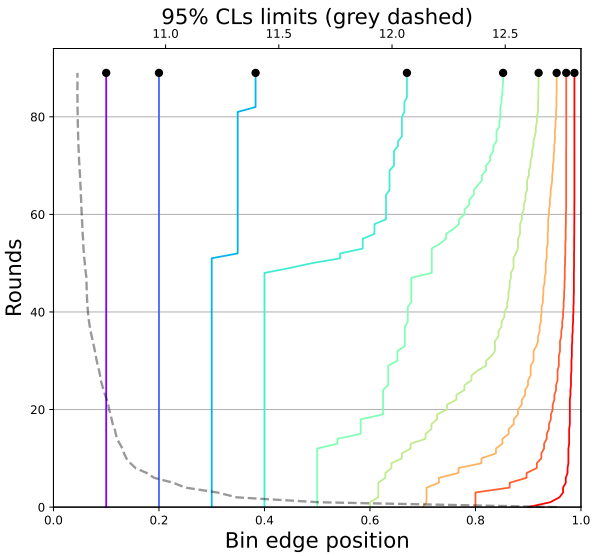
Figures 28 and 29 show optimization histories of toys for different touch-rates where statistical constraints are enforced. The change in binning focuses again first on the significant bins, but becomes less extreme, still fulfilling the constraints.



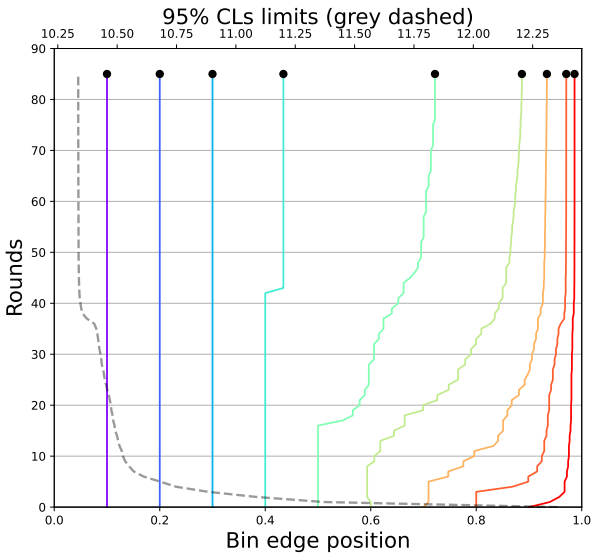
(a) Touch-rate 5, original data

(b) Touch-rate 10, toy number 146

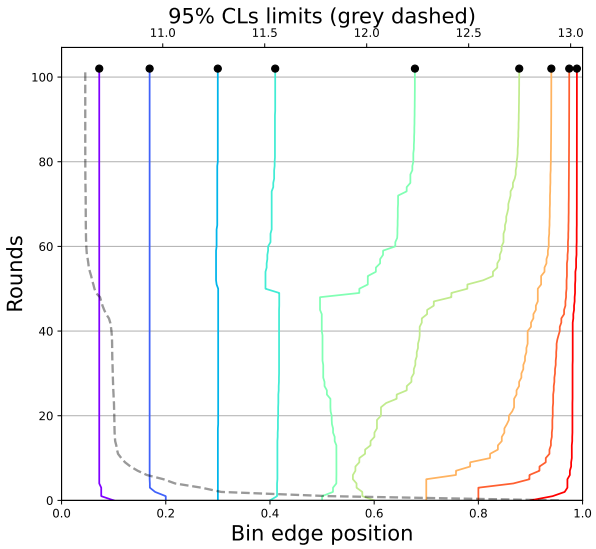
Figure 28: Optimization of the original data and toys for different touch-rates with constraints enforced. The grey dashed line refers to the 95% CL_s limit on the upper x-axis. The colored lines represent the bin edges on the lower x-axis.



(a) Touch-rate 15, toy number 11



(b) Touch-rate 20, toy number 66



(c) Touch-rate 25, toy number 10

Figure 29: Optimization of the original data and toys for different touch-rates with constraints enforced. The grey dashed line refers to the 95% CL_s limit on the upper x-axis. The colored lines represent the bin edges on the lower x-axis.

A lot of optimizations fail in the first round when constraints are enforced. This includes also the original data for most touch-rates.
To account for that, more toys are produced to get similar amounts for the evaluation plots. Thus 2500 toys are used for the constrained spline optimization, of which a fraction is optimized successfully. The limit distributions for the different touch-rates are shown in Figure 30. The amount of failed optimizations grows with the touch rate. For a touch-rate of five 60% of the toys and also the original data could be optimized successfully. For the highest touch-rate this drops to only 15% of the toys and not the original data anymore. The mean optimized limit is closer to the mean start limit for the smaller touch-rates, with a difference of approximately 1.3, than the unconstrained one, but for a higher touch-rate the difference grows to also approximately two. The standard deviations are similar to the unconstrained ones.

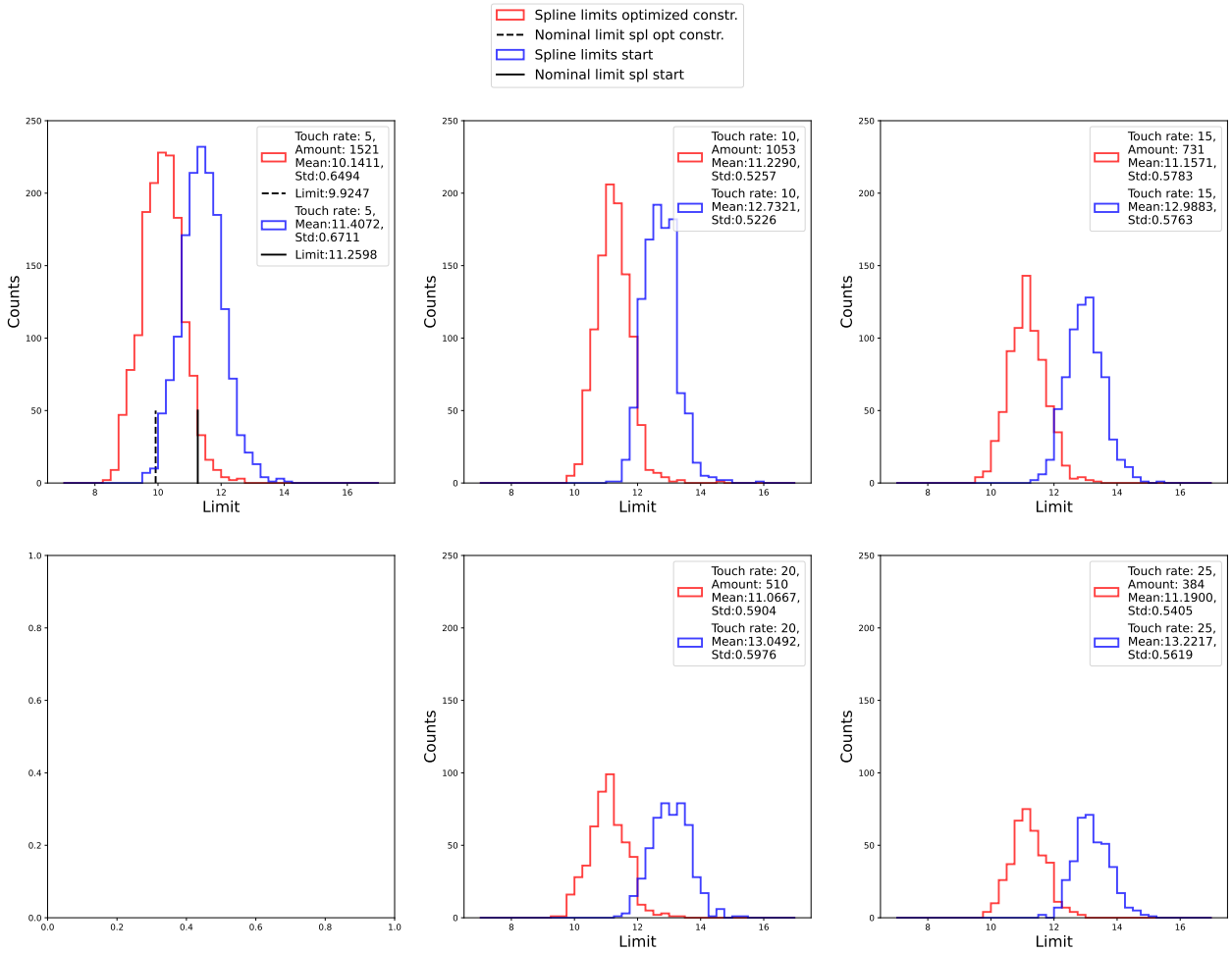


Figure 30: Spline limit distributions with constraints for different touch-rates. The mean values (mean), standard deviations (std) and the amount of data points that succeeded the optimization (toys plus original) are shown in the legends.

These results may suffer from a survivorship bias. To see the maximum possible bias, the toys for which the optimization was successful for all touch-rates are plotted again without the rest. These toys are further referred to as touch-rate passing toys. The 210 toys that are touch-rate passing are plotted in Figure 31.
The changes compared to the distributions of all possible toys are minor, indicating that a possible survivorship bias is small.

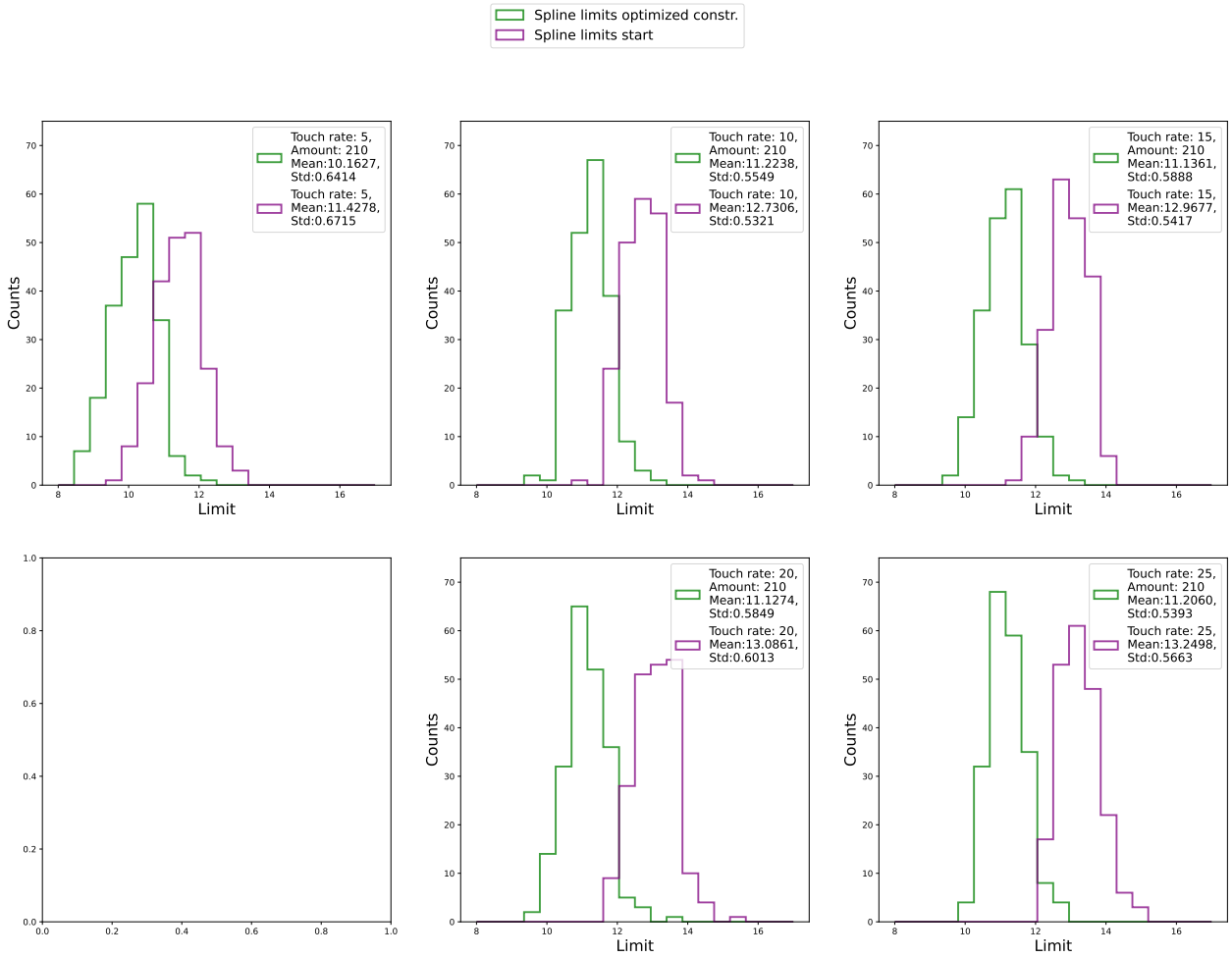
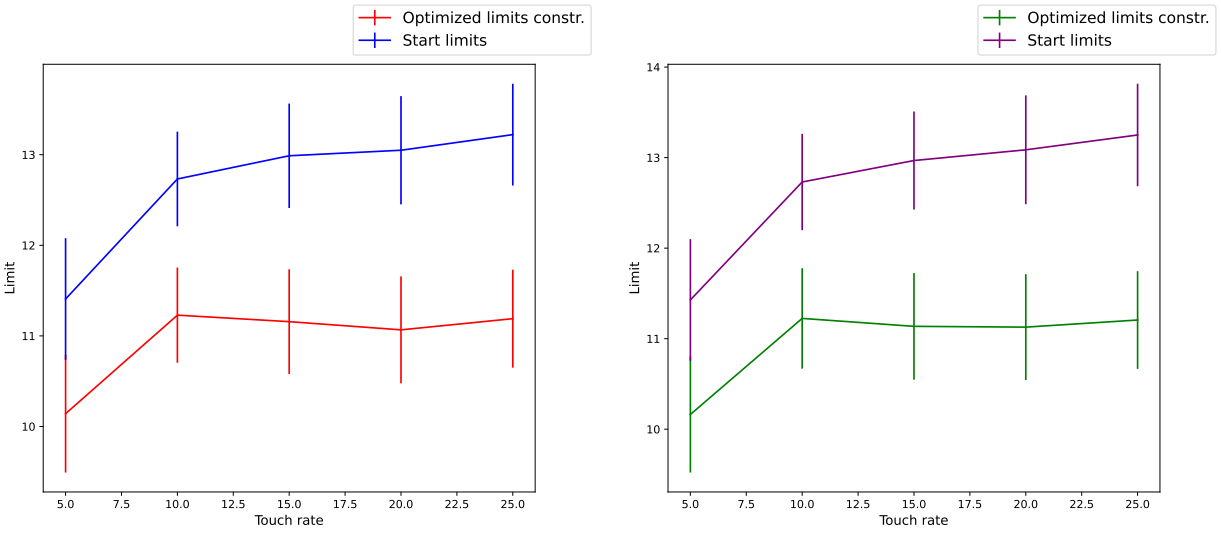


Figure 31: Spline limit distributions with constraints for all touch-rates passing toys. The mean values (mean), standard deviations (std) and the amount of data points (210 toys) are shown in the legends.

The plots of mean limit for each touch rate in Figure 32 reflect the similarities again. Compared to the unconstrained optimization the convergence of the mean limit regarding the touch-rate seems more definitive.



(a) All toys that succeeded for each touch-rate.

(b) Only all touch-rates passing toys.

Figure 32: Shown is the mean limit as a function of the touch-rate, with constraints in the optimization, for all toys that succeeded for each touch-rate in 32(a) and for all 210 touch-rate passing toys in 32(b). The error bars denote the width of the limit distributions.

Figures 33 and 34 show the distributions of differences between limits before (equidistant binning) and after optimization. This is shown for all toys and the 210 touch-rate passing toys respectively. The widths are relative to the average improvements approximately 11% for the touch-rate of 5, 18% for the touch-rate of 10, 22% for the touch-rate of 15, 16% for the touch-rate of 20 and 17% for the touch-rate of 25.

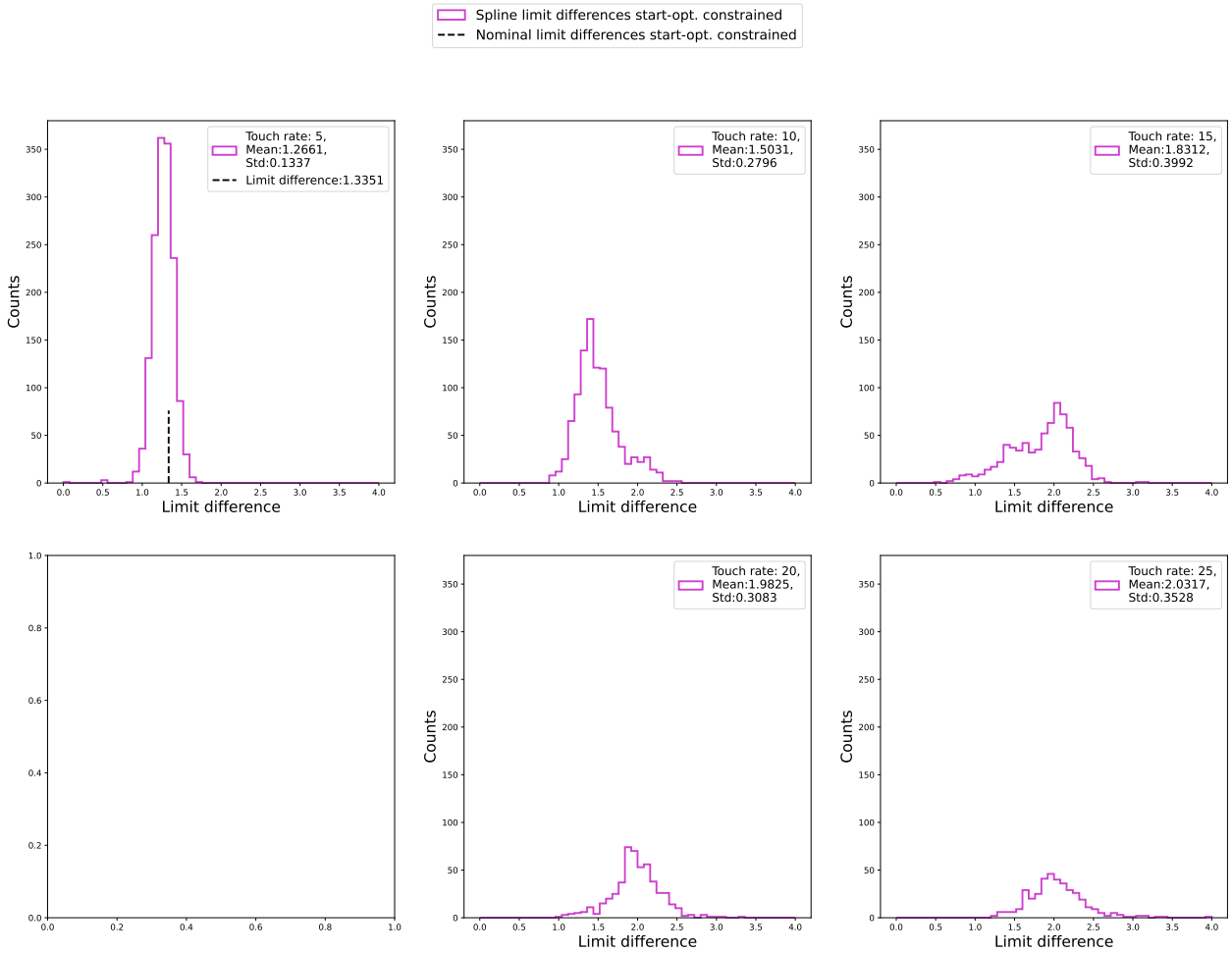


Figure 33: Distributions of differences between limit before (equidistant binning) and after optimization with constraints enforced for different touch-rates. The black line indicates the original data. Shown are all toys that succeeded for each touch-rate. The mean values (mean) and standard deviations (std) are shown in the legends.

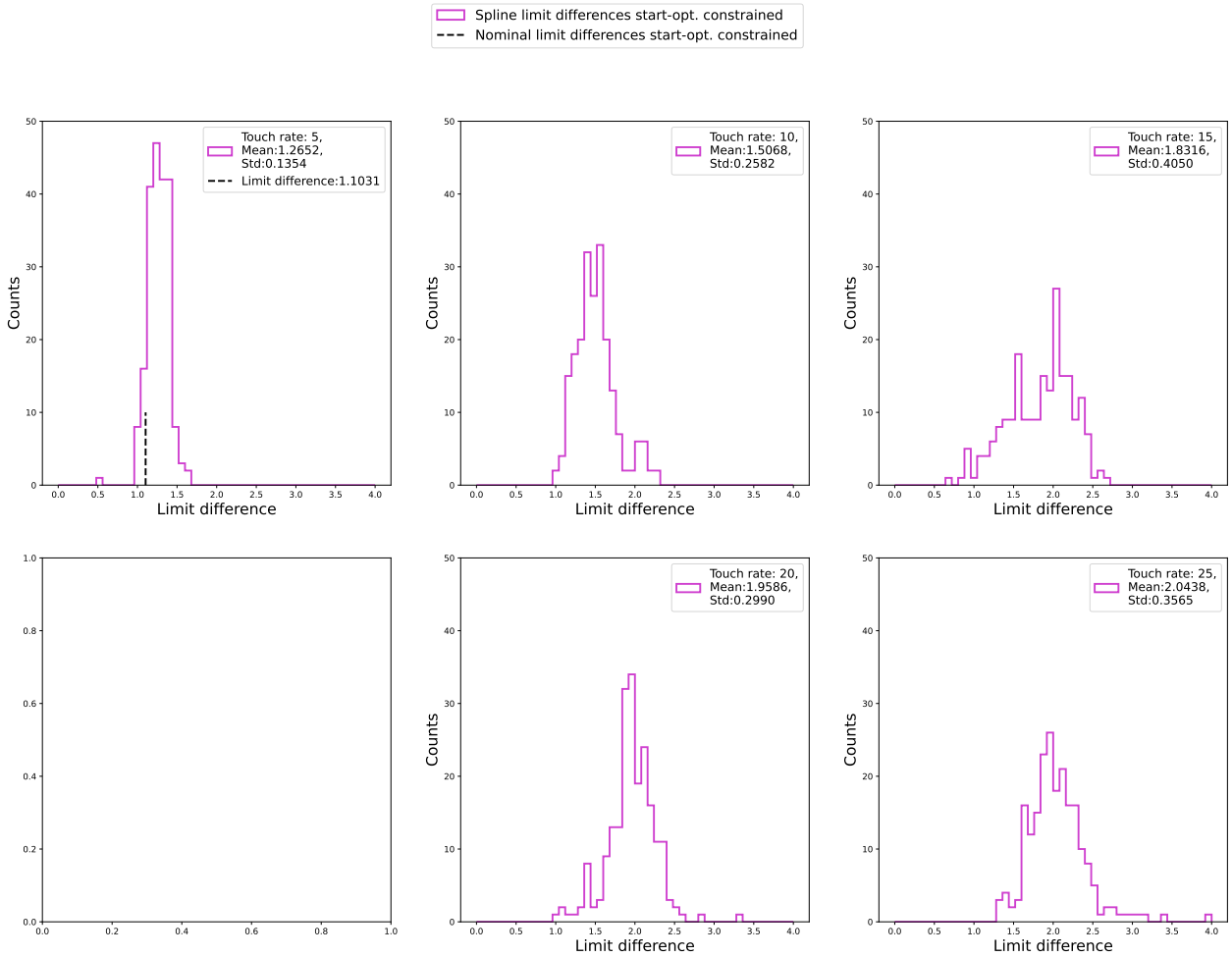


Figure 34: Distributions of differences between limit before (equidistant binning) and after optimization with constraints enforced for different touch-rates. The black line is a mistake in plotting and indicates some toy, not the original data. Shown are only the 210 touch-rate passing toys. The mean values (mean) and standard deviations (std) are shown in the legends.

Comparing the widths of the limit differences before and after optimization, the spline optimizations show smaller ones, especially without constraints and small touch-rates, than the widths of the limit differences for the mini-bin optimizations. Without constraints enforced, higher touch-rates lead to larger widths in these distributions. This is expected, as the splines represent the mini-bins more precise. When enforcing constraints the width is largest for the touch-rate of 15. This could be coincidental and higher touch-rates than 25 should surpass that.

In Figures 35 and 36 the 2-D plots of the mini-bin transformations are depicted. The trend of the transformed limits for the different touch-rates is similar to the transformed limits where no constraints are enforced in the optimization. The average improvement ranges for the touch-rate of 5 with about 1.3 to approximately 2 for the touch-rate of 25 for all toys and also the touch-rate passing toys. This are significant improvements compared to the average values of 11.4 for the touch-rate of 5 and up to 13.2 for the touch-rate of 25 before optimization.

Compared to the constrained mini-bin optimization mean limit of 11.97 ± 0.54 the transformed mean limit is slightly larger (+0.30) for the smallest touch-rate and very slightly below (-0.09) for the largest touch-rate. The transformations also approach the identity for higher touch-rates and more limits stay below the identity line. The transformed distributions for the touch-rate passing toys has very similar values. A survivorship bias should show differences, thus it is unlikely that the bias exists.

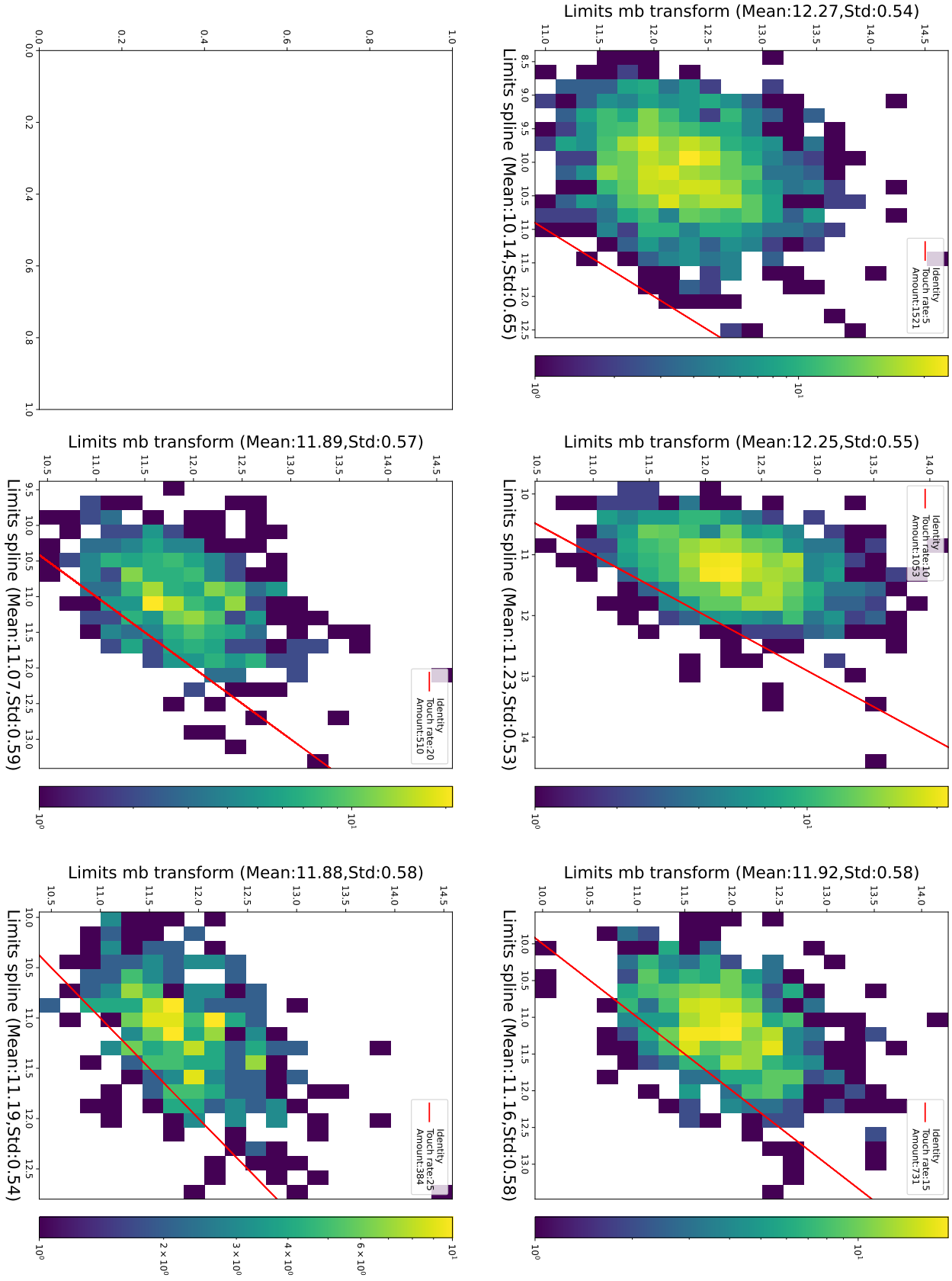


Figure 35: The transformation of optimized spline limits to limits based on mini-bin yields with constraints is shown as a 2-D histogram. The red line shows the identity transformation. The mean values (mean) and standard deviations (std) are shown at the axes labels.

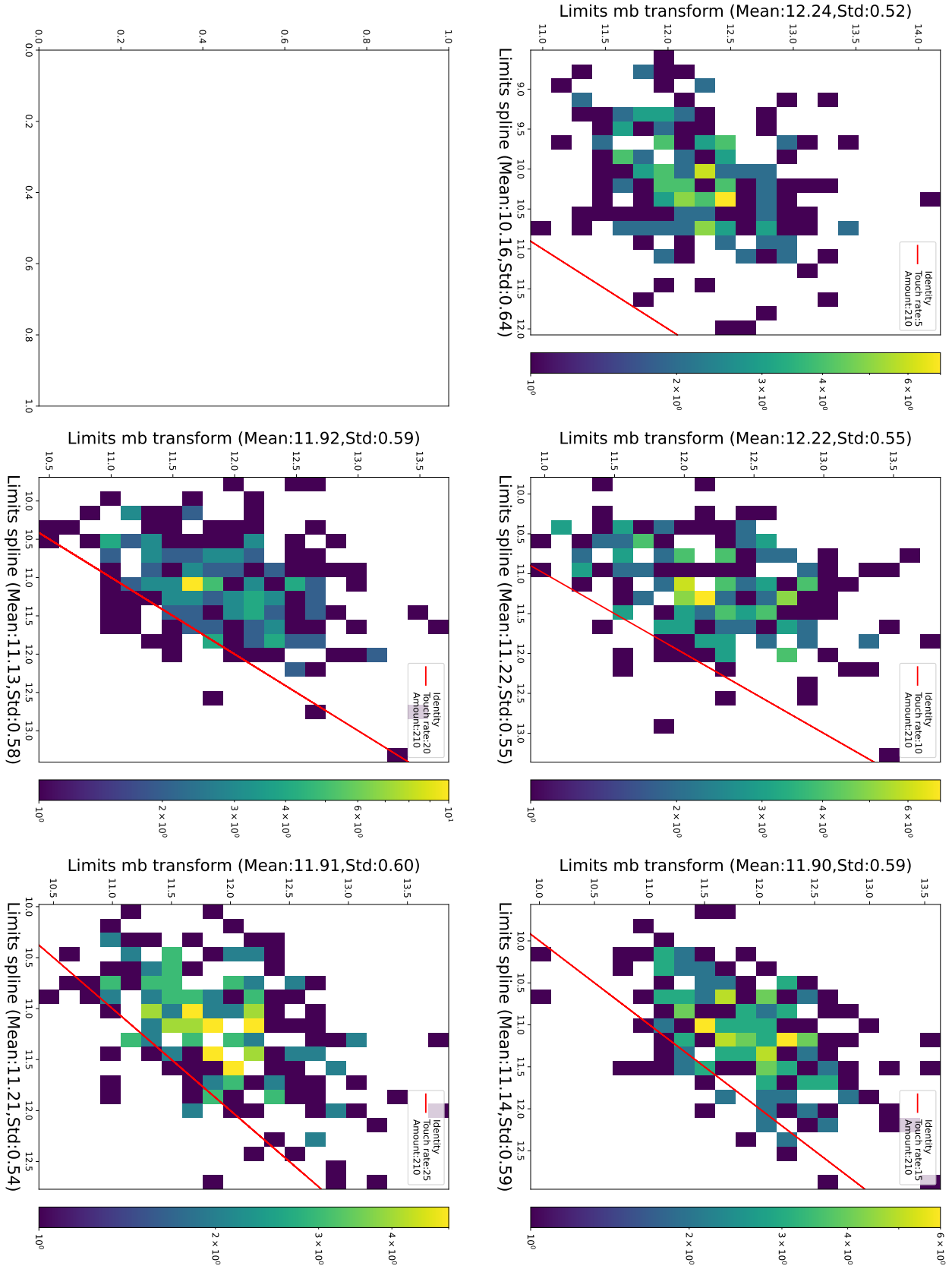


Figure 36: The transformation of optimized spline limits to limits based on mini-bin yields with constraints, for the 210 touch-rate passing toys, is shown as a 2-D histogram. The red line shows the identity transformation. The mean values (mean) and standard deviations (std) are shown at the axes labels.

8 Summary and Outlook

In this work a novel methodology and setup are explored to optimize the binning of a discriminator variable regarding the upper limit.
This is part of a larger effort aiming for a fully differentiable analysis chain that can be optimized from end to end in all of its aspects. The immediate use however is, that the details of the binning of discriminators is known to have a significant impact on the results of current statistical inference algorithms, that fully rely on binned distributions of classifiers for signal and background. The approach chosen here makes use of new tools that provide a differentiable limit as function of bin contents. This is used to formulate an optimization of the limit with respect to the binning.
Two methods with different approaches are examined. The mini-bin method optimizes the binning directly on the output of the DNN distributions. And the spline method optimizes the binning on analytic representations of these distributions. Both are studied including also statistical constraints on the size of the background samples.

It is shown, that the general setup, the methodology and the two methods are able to achieve the optimizations on the upper limit without assumptions or intermediate optimization steps. However, there are caveats that should be noted.
The technical setup does not allow differentiation of a statistical model with more than one systematic or MC uncertainties, but the tools to do so are still in development.
The mini-bin method works, but is not able to over-optimize and is not safe against intermediate "bad" steps for the limit. The splining method is not limited by these aspects, but needs a careful hand in the creation of the splines to "correctly" represent the underlying distributions. Also the optimization with statistical constraints can be complicated to be successful, using a standard optimizer.
Important possible improvements are therefore the construction method of the splines and the optimization with constraints.

For application of the spline method it needs to be decided on a touch-rate by evaluating the limit difference, chi-squared and other distributions shown in this thesis for a set of variations on the original data. This can be done on a different number of fixed bins to optimize the classifier histograms in both aspects of number of bins and bin positions.

Looking ahead the methodology can be extended to optimize the upper limit or any other summary statistic in all categories of a multiclass problem at the same time, without assumptions or intermediate optimizations.

A Appendix

A.1 Data distributions with Spline derivatives

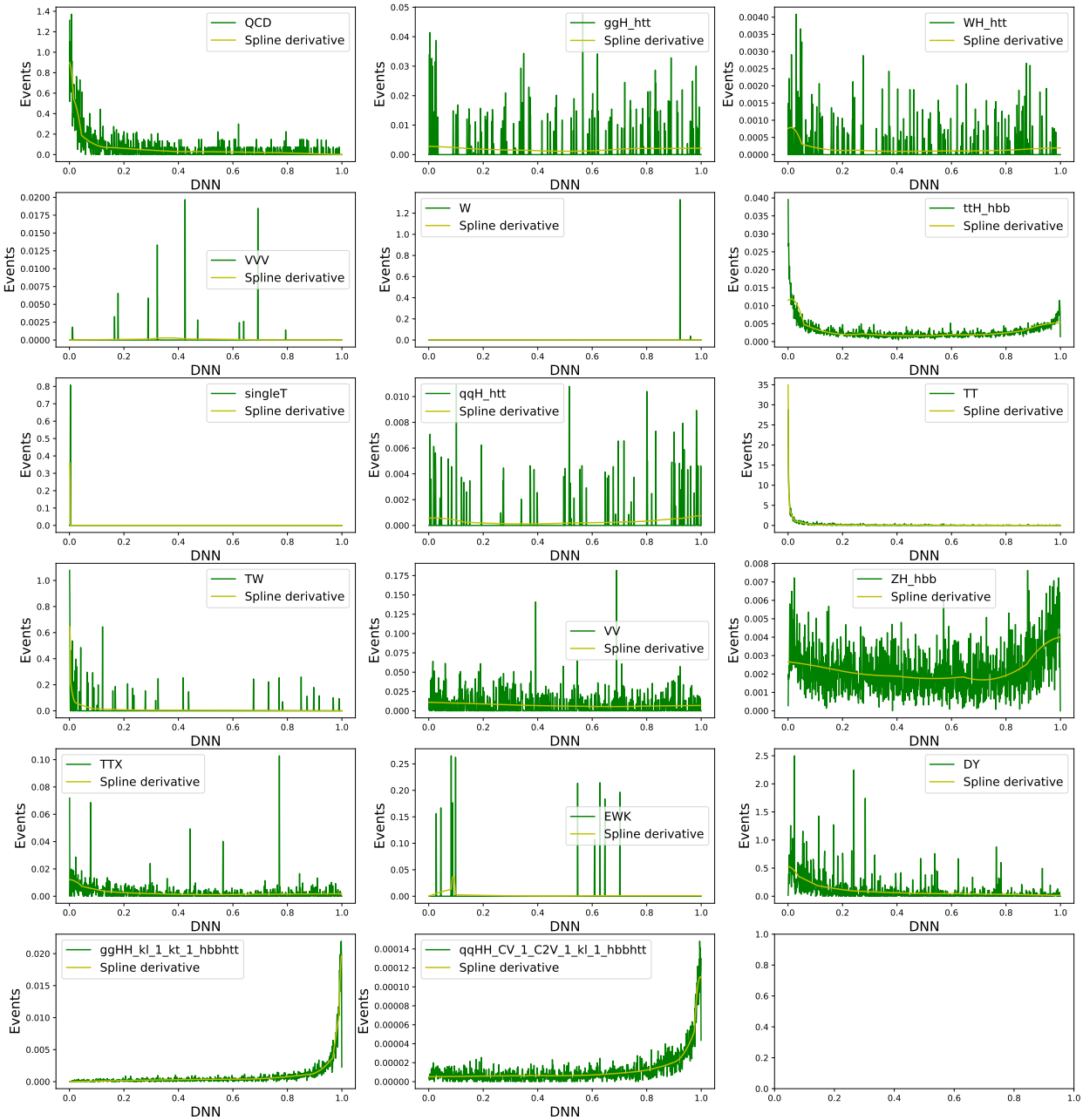


Figure 37: Scaled derivatives of splines for each process and a touch-rate of 5 for the original data.

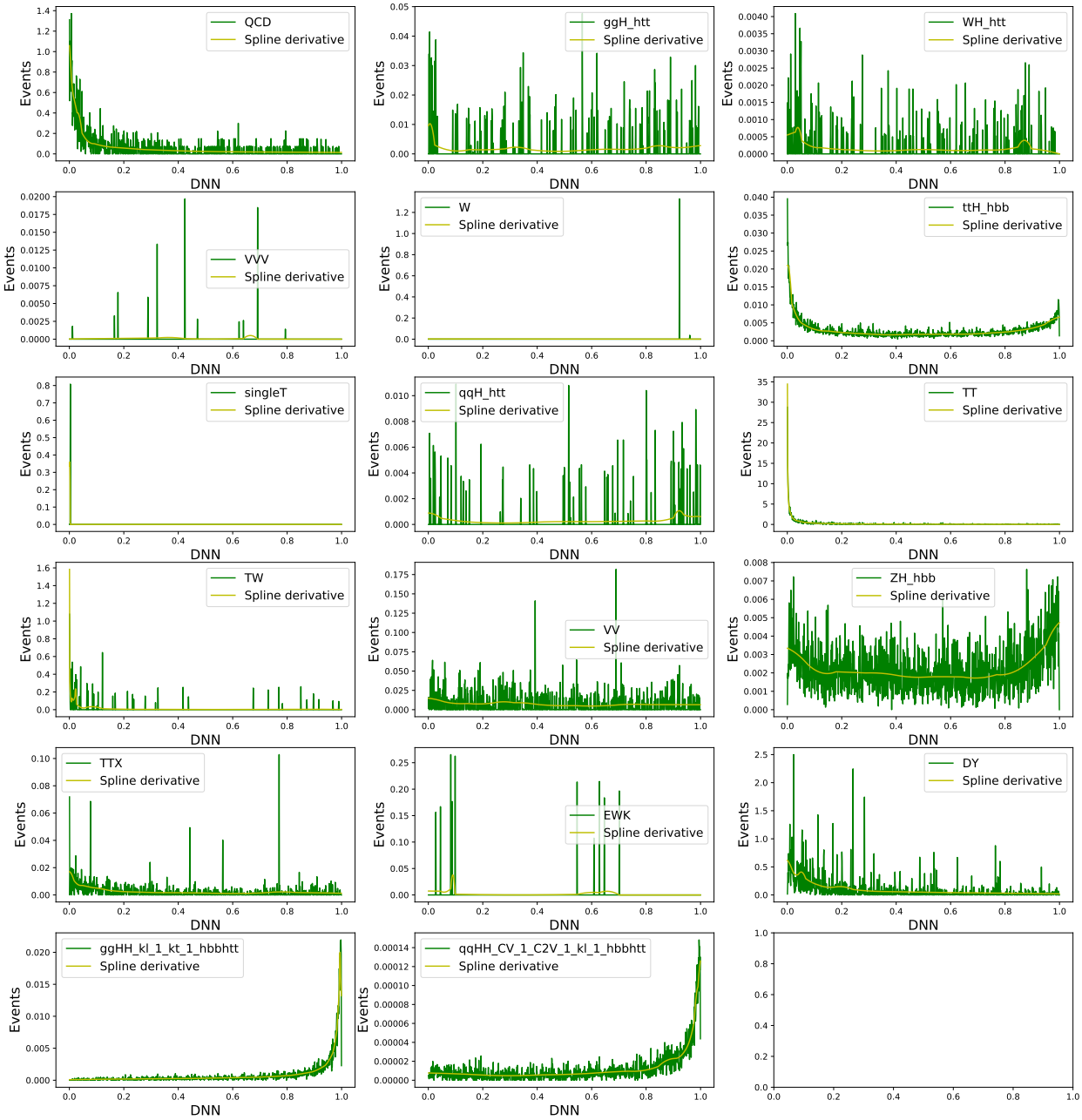


Figure 38: Scaled derivatives of splines for each process and a touch-rate of 10 for the original data.

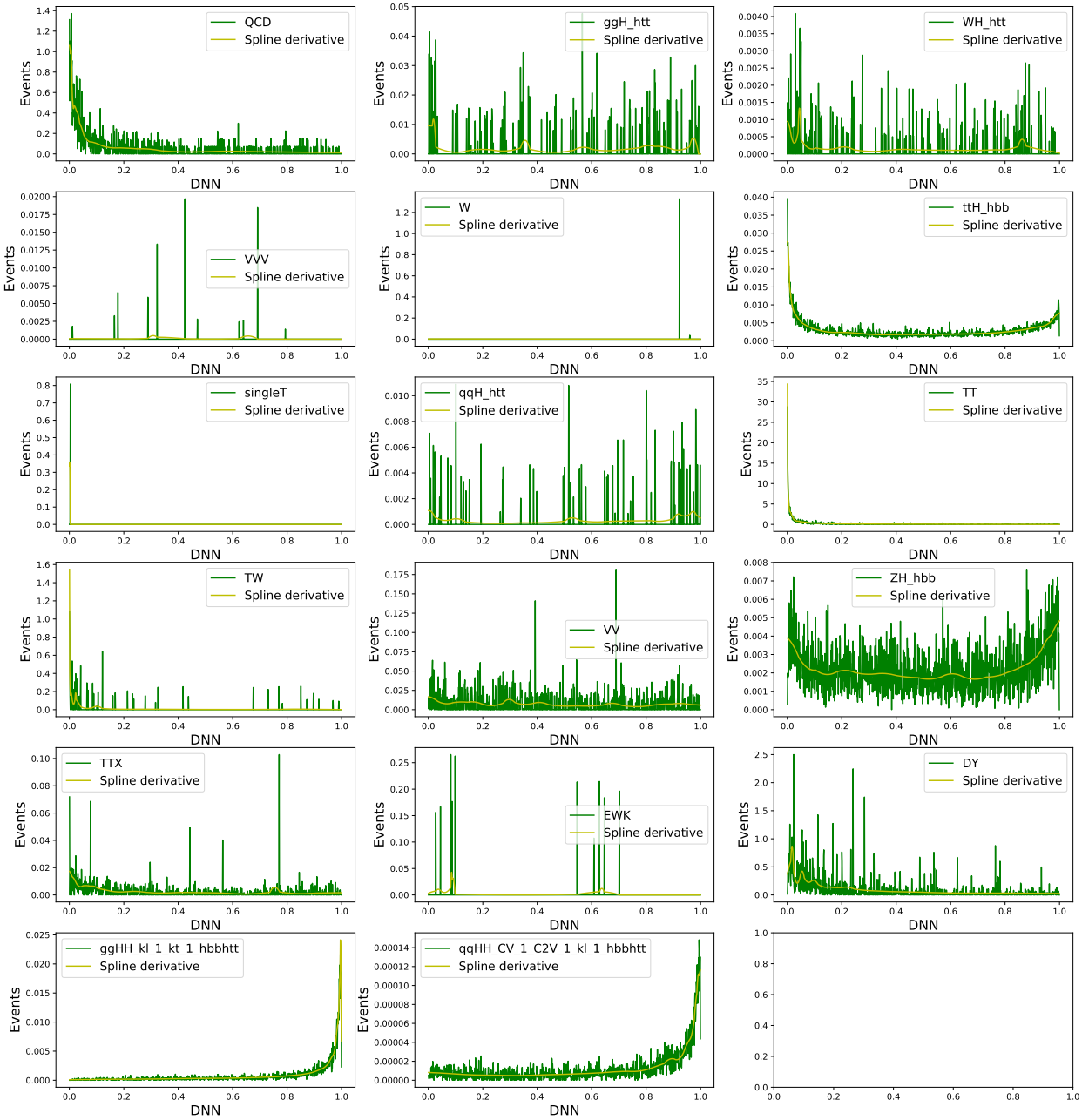


Figure 39: Scaled derivatives of splines for each process and a touch-rate of 15 for the original data.

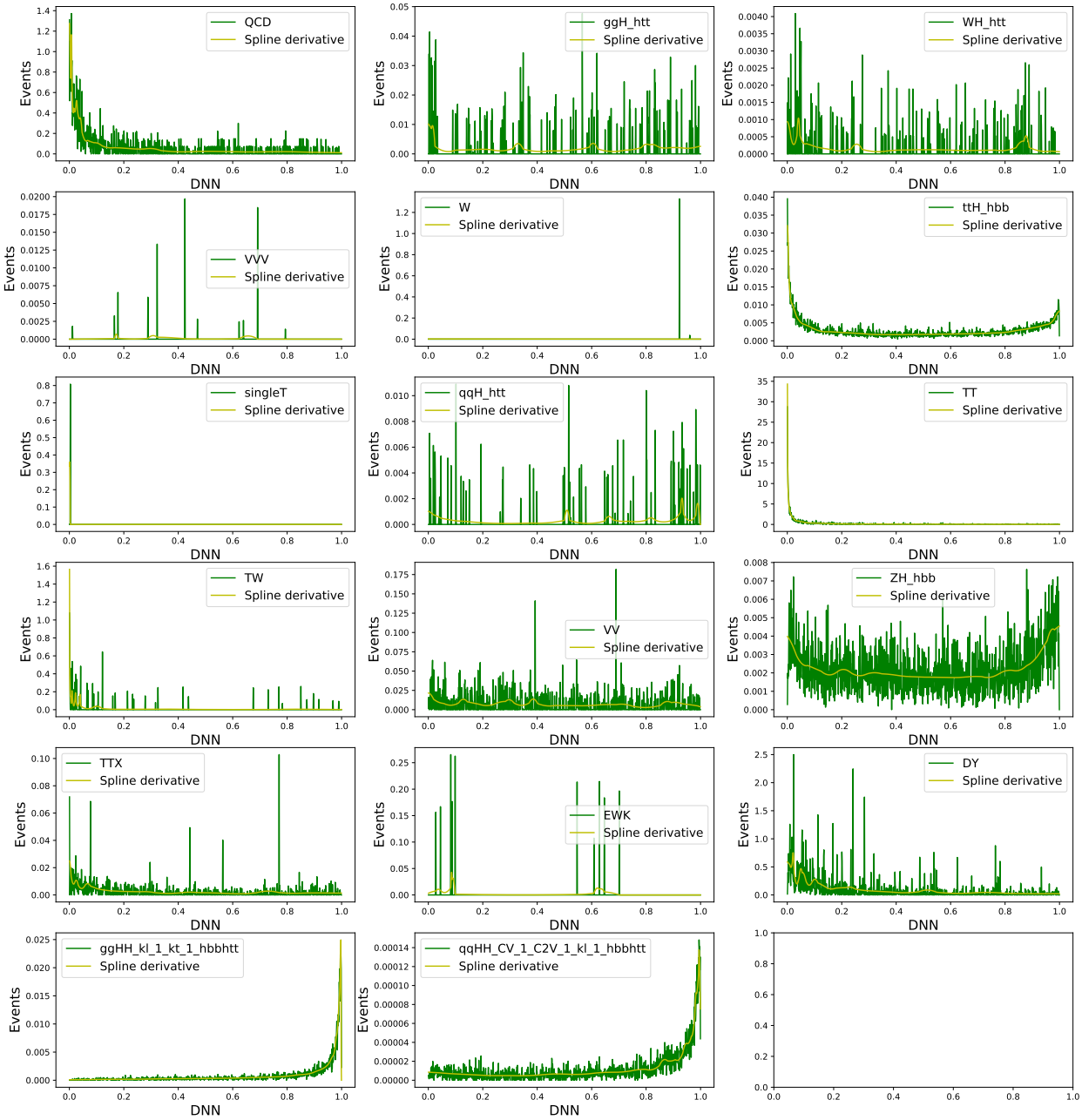


Figure 40: Scaled derivatives of splines for each process and a touch-rate of 20 for the original data.

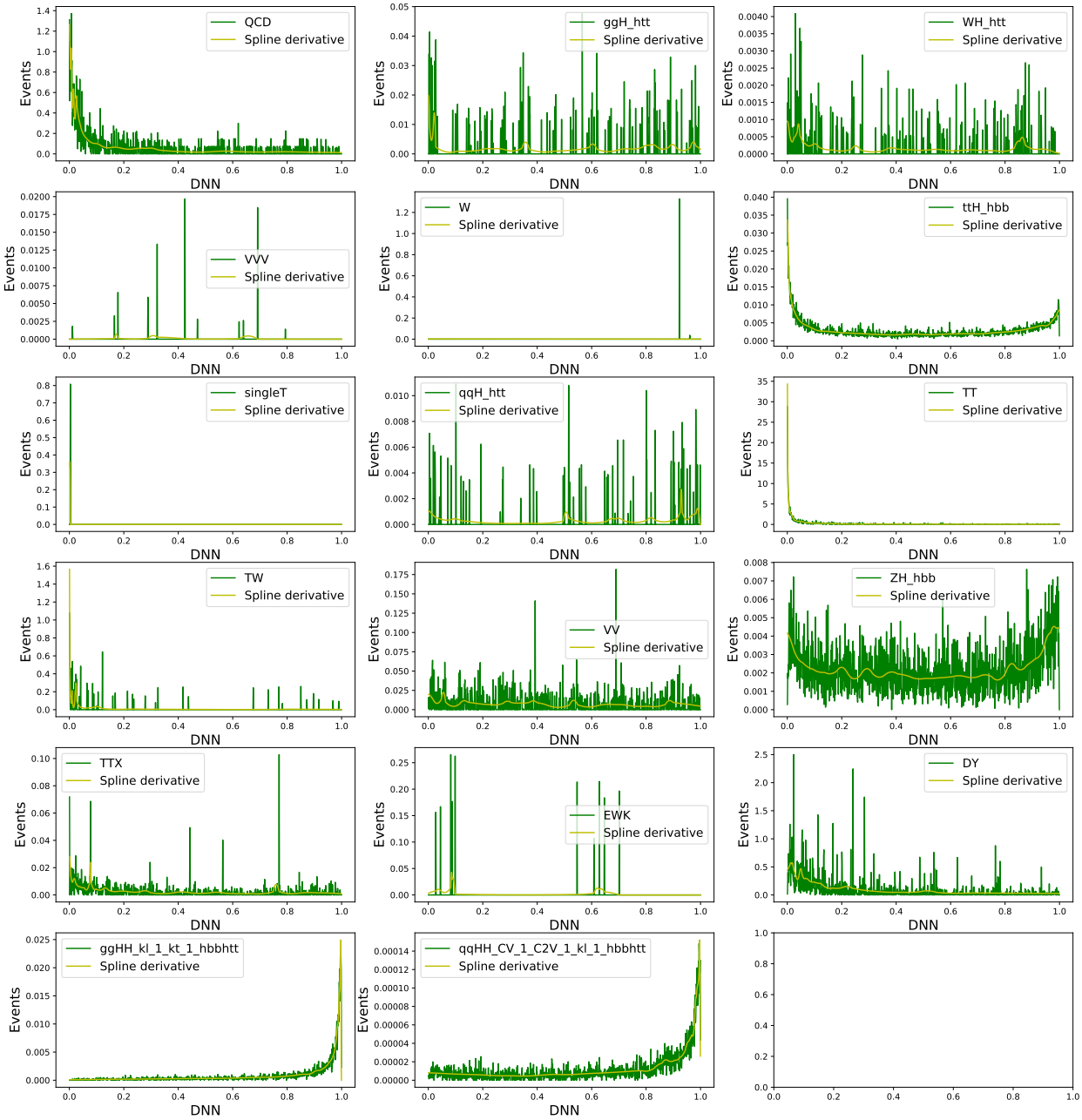


Figure 41: Scaled derivatives of splines for each process and a touch-rate of 25 for the original data.

A.2 Splines of Cumulative distributions

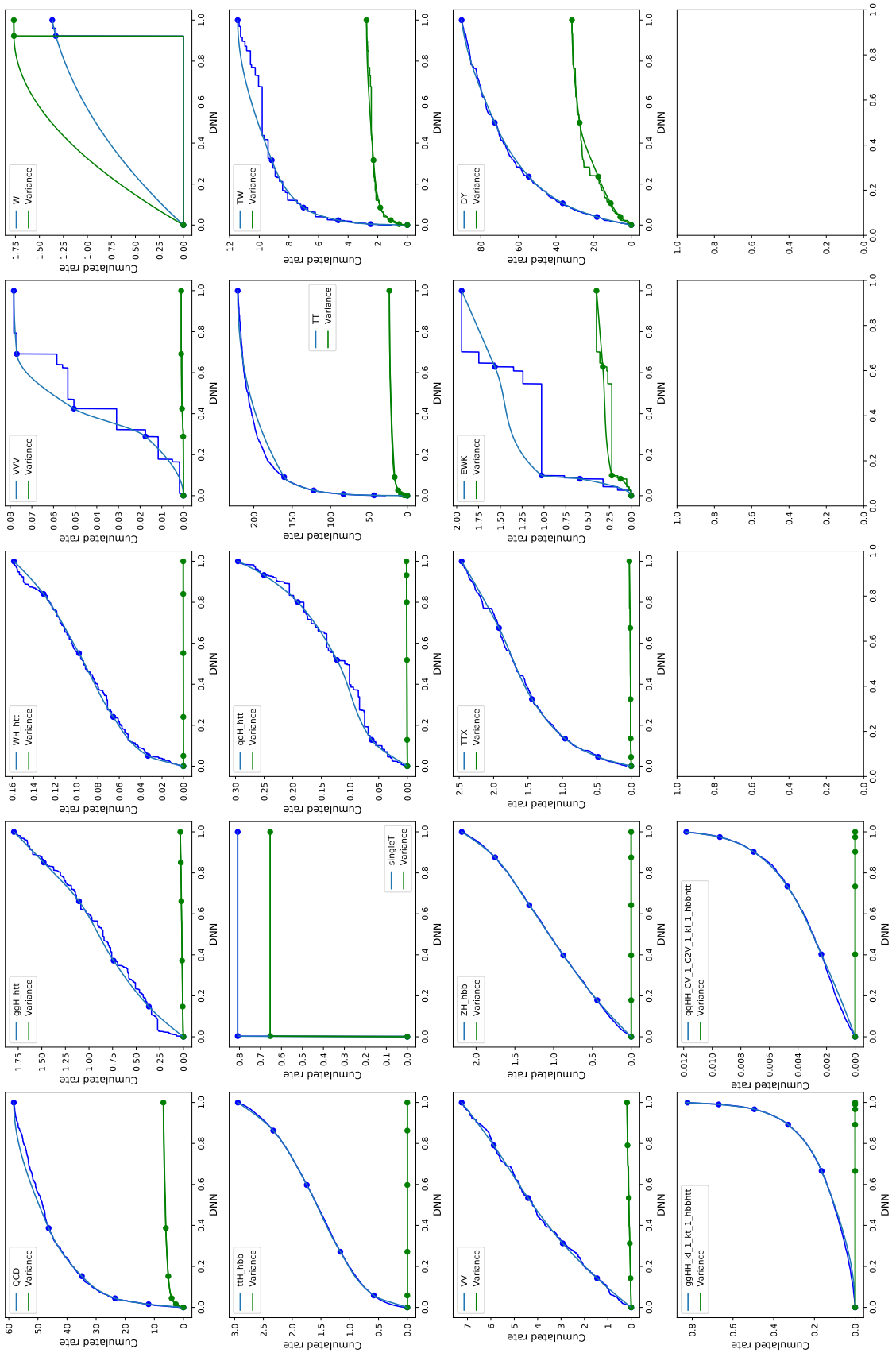


Figure 42: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 5. Shown are for the original data the results for all signal and background processes (see legends).

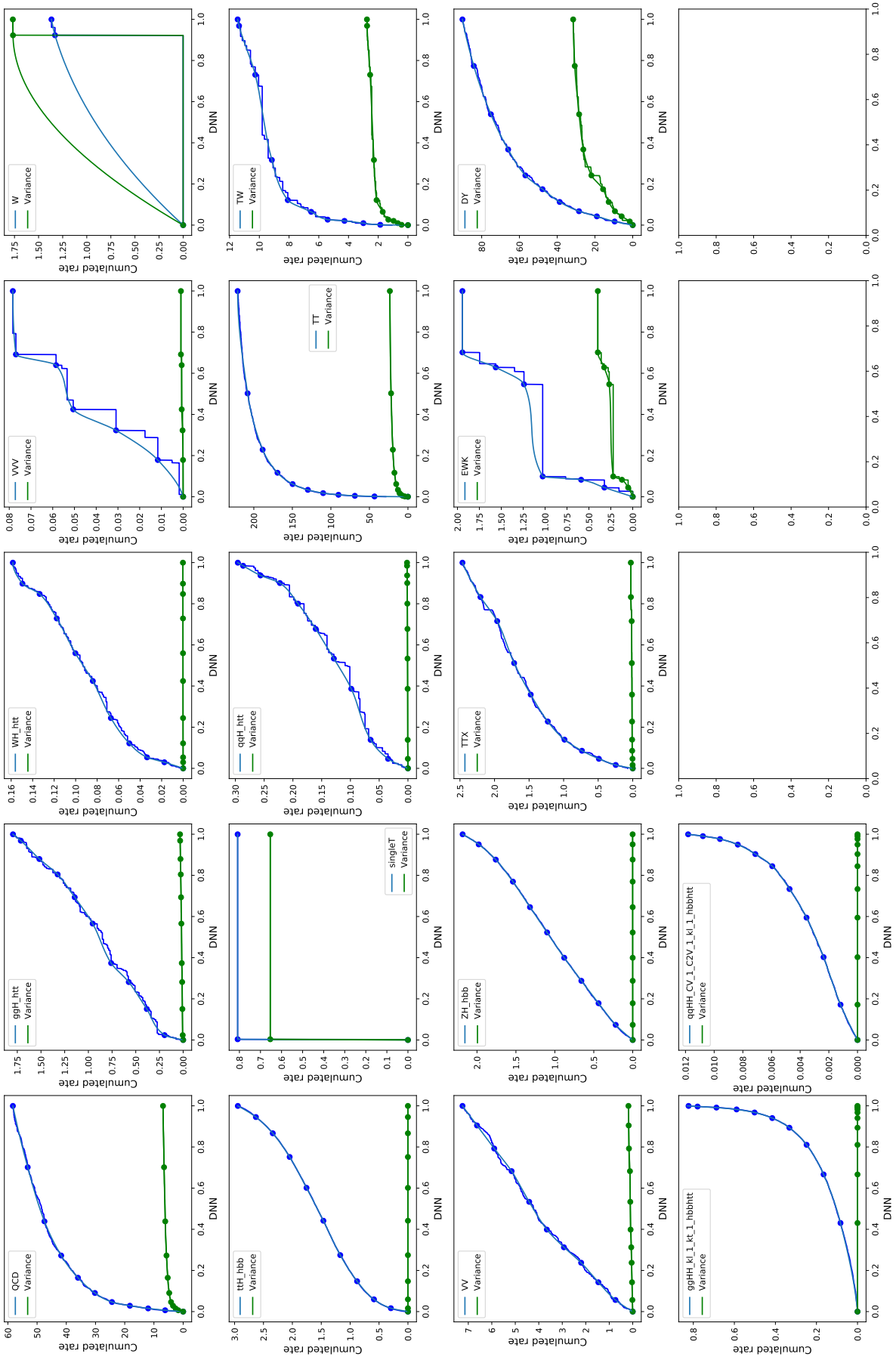


Figure 43: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 10. Shown are for the original data the results for all signal and background processes (see legends).

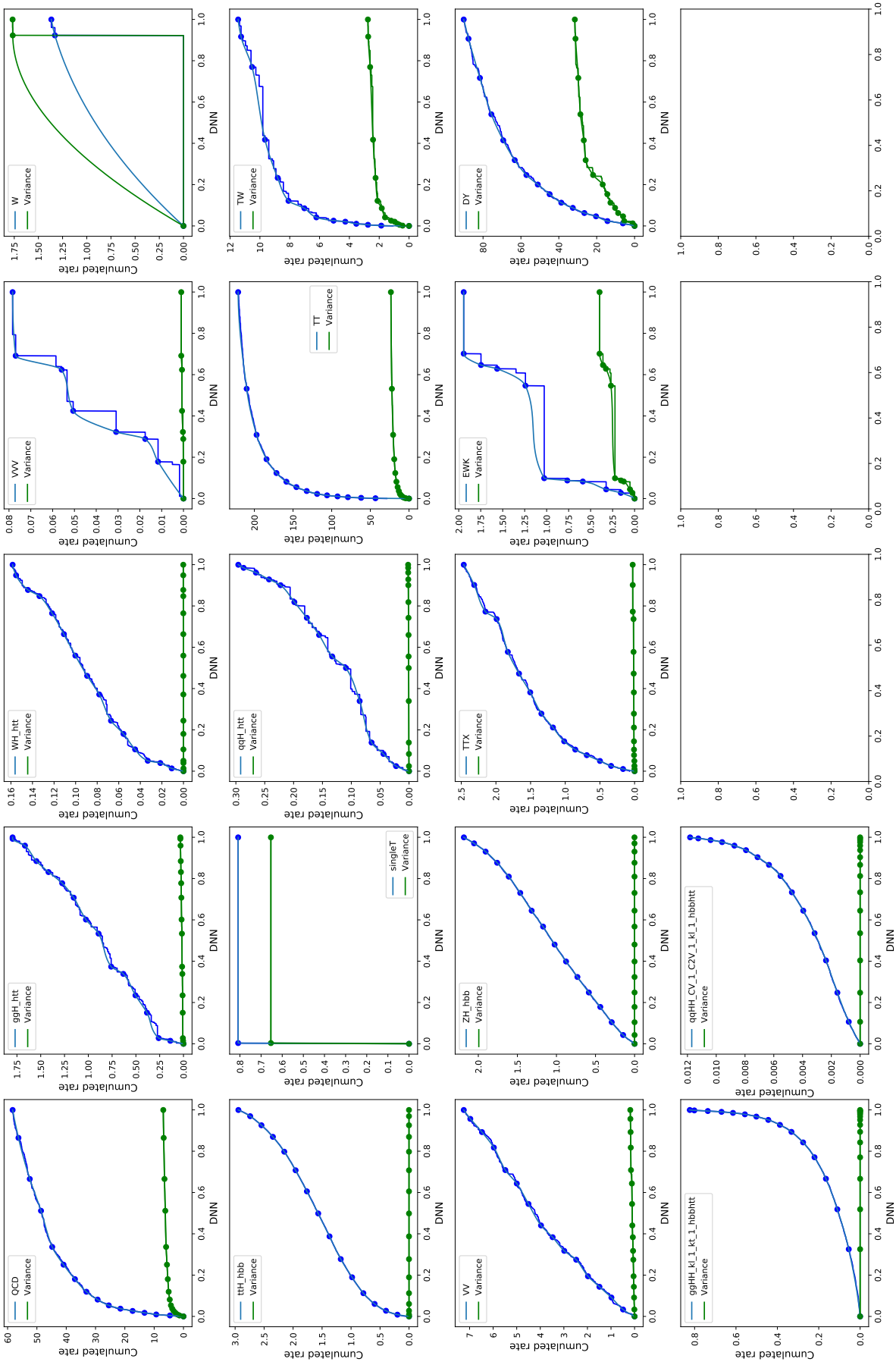


Figure 44: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 15. Shown are for the original data the results for all signal and background processes (see legends).

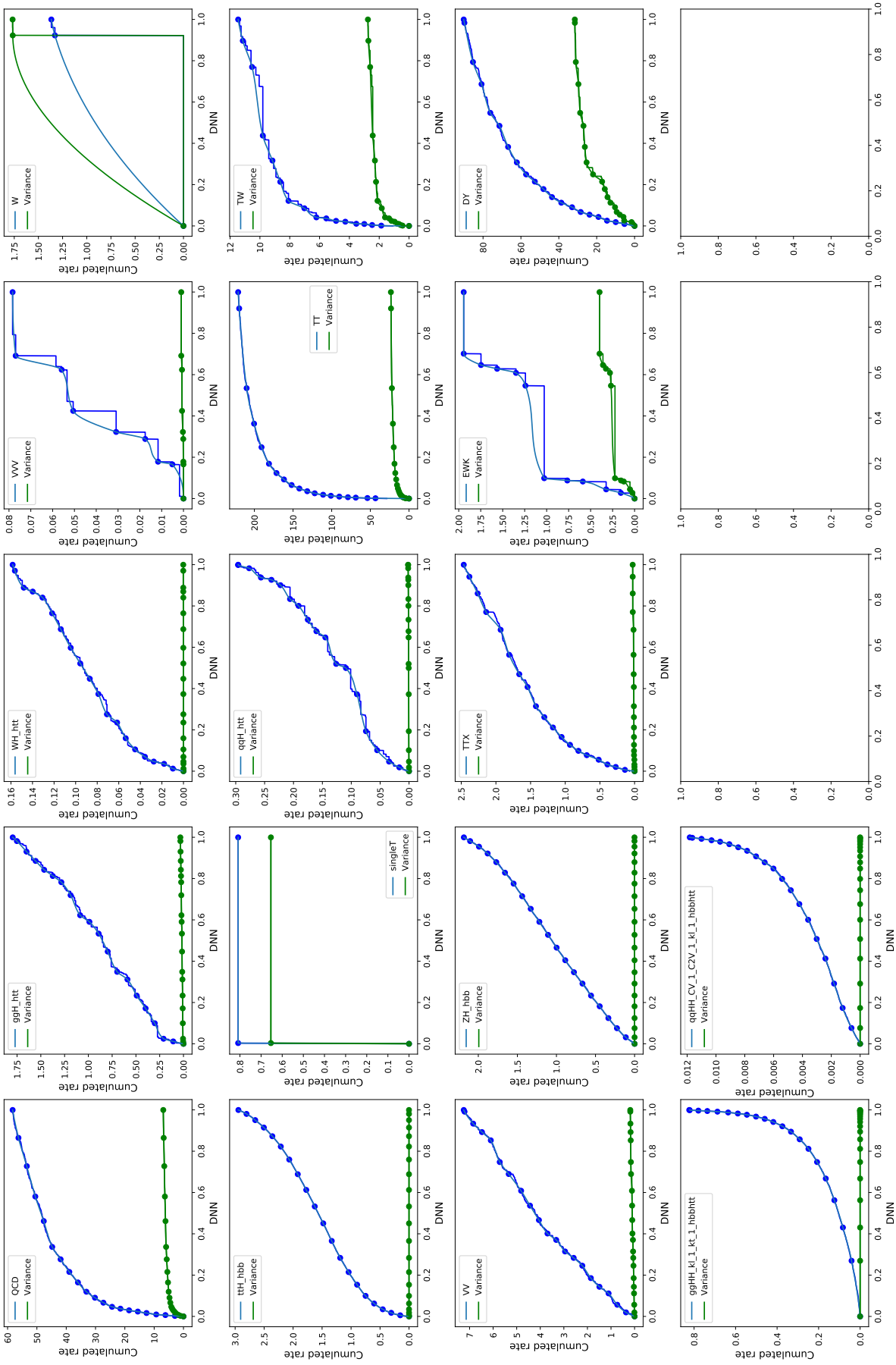


Figure 45: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 20. Shown are for the original data the results for all signal and background processes (see legends).

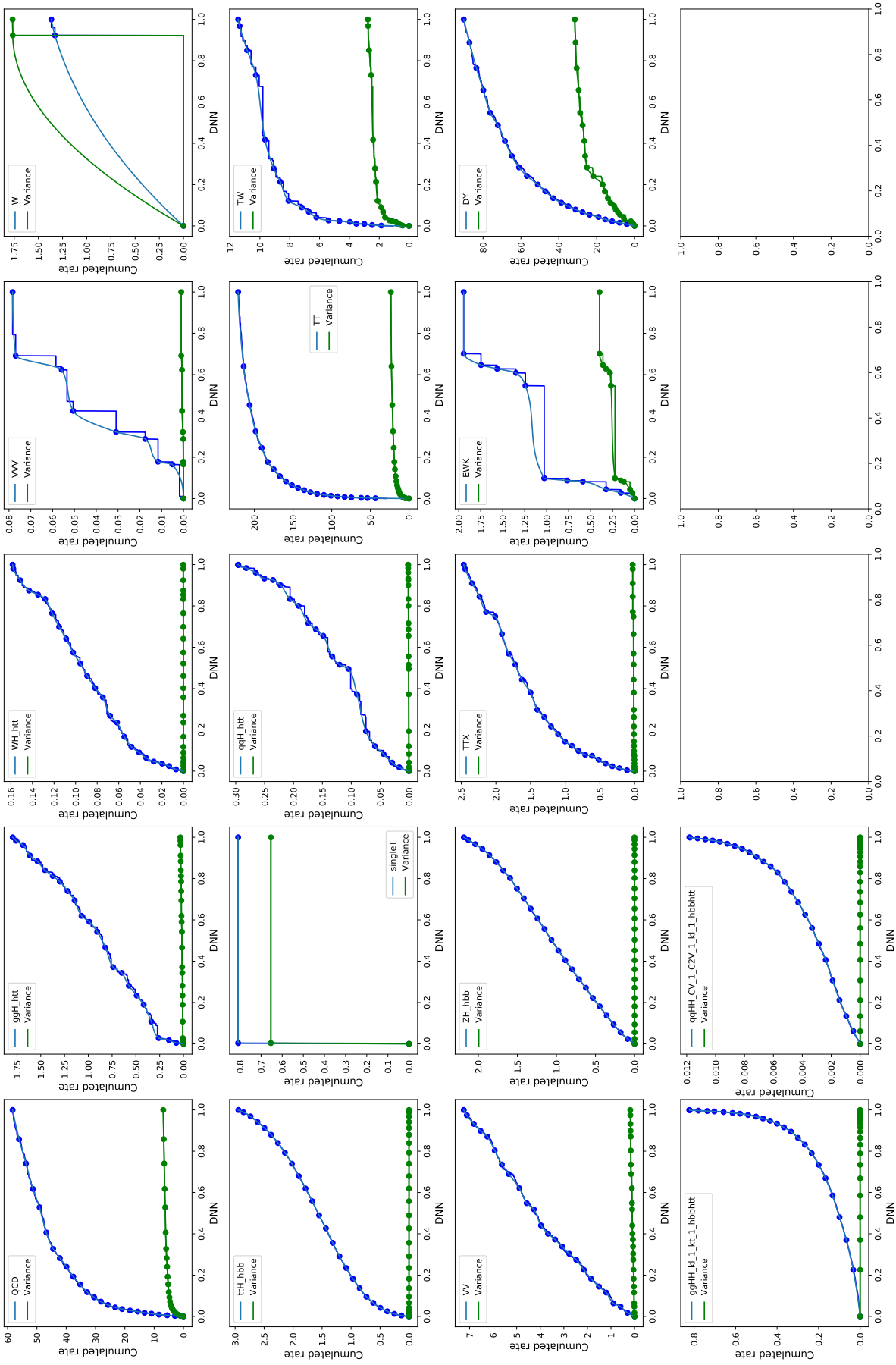


Figure 46: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 25. Shown are for the original data the results for all signal and background processes (see legends).

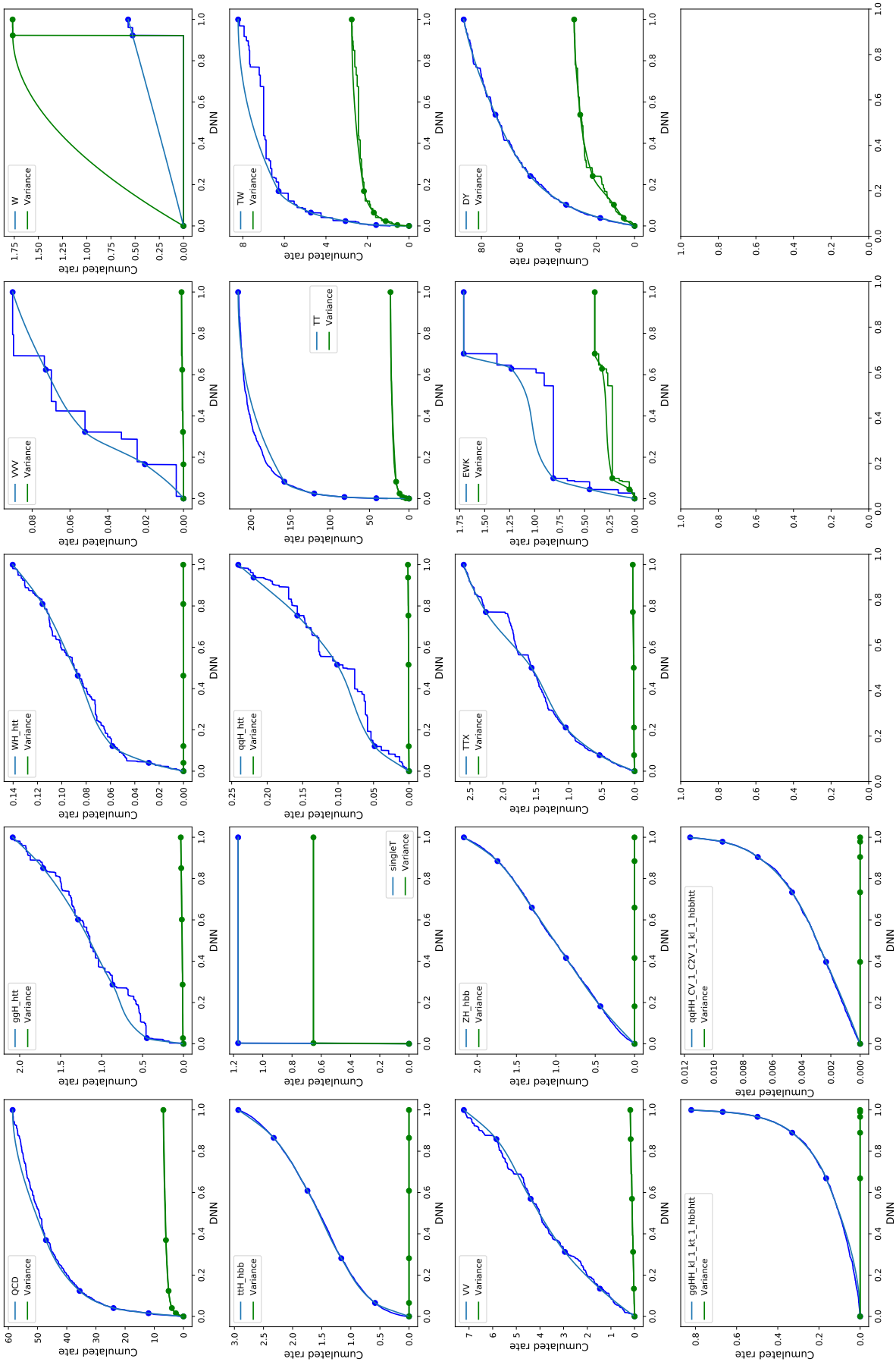


Figure 47: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 5. Shown are for the toy data of toy number 1 the results for all signal and background processes (see legends).

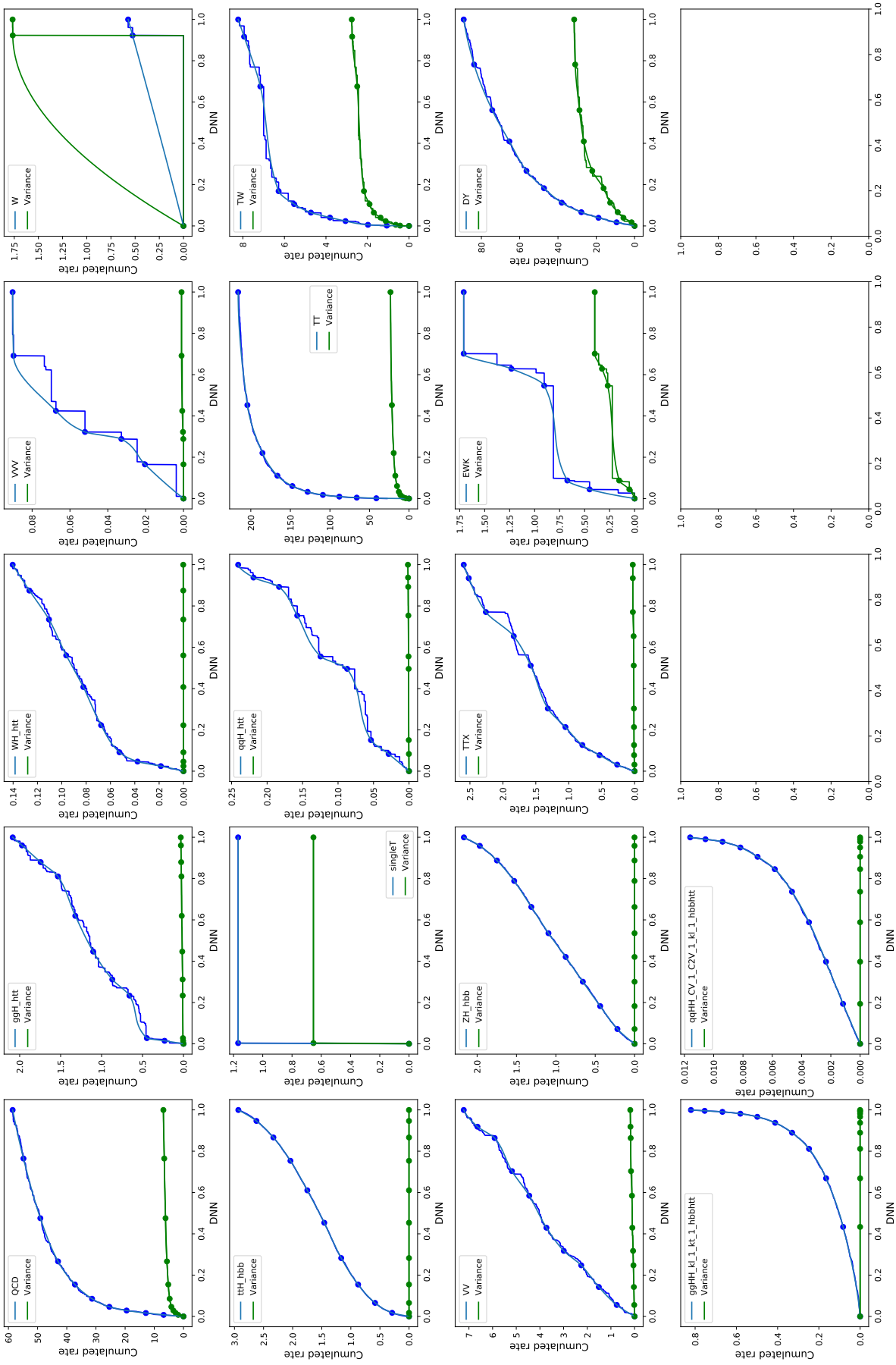


Figure 48: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 10. Shown are for the toy data of toy number 1 the results for all signal and background processes (see legends).

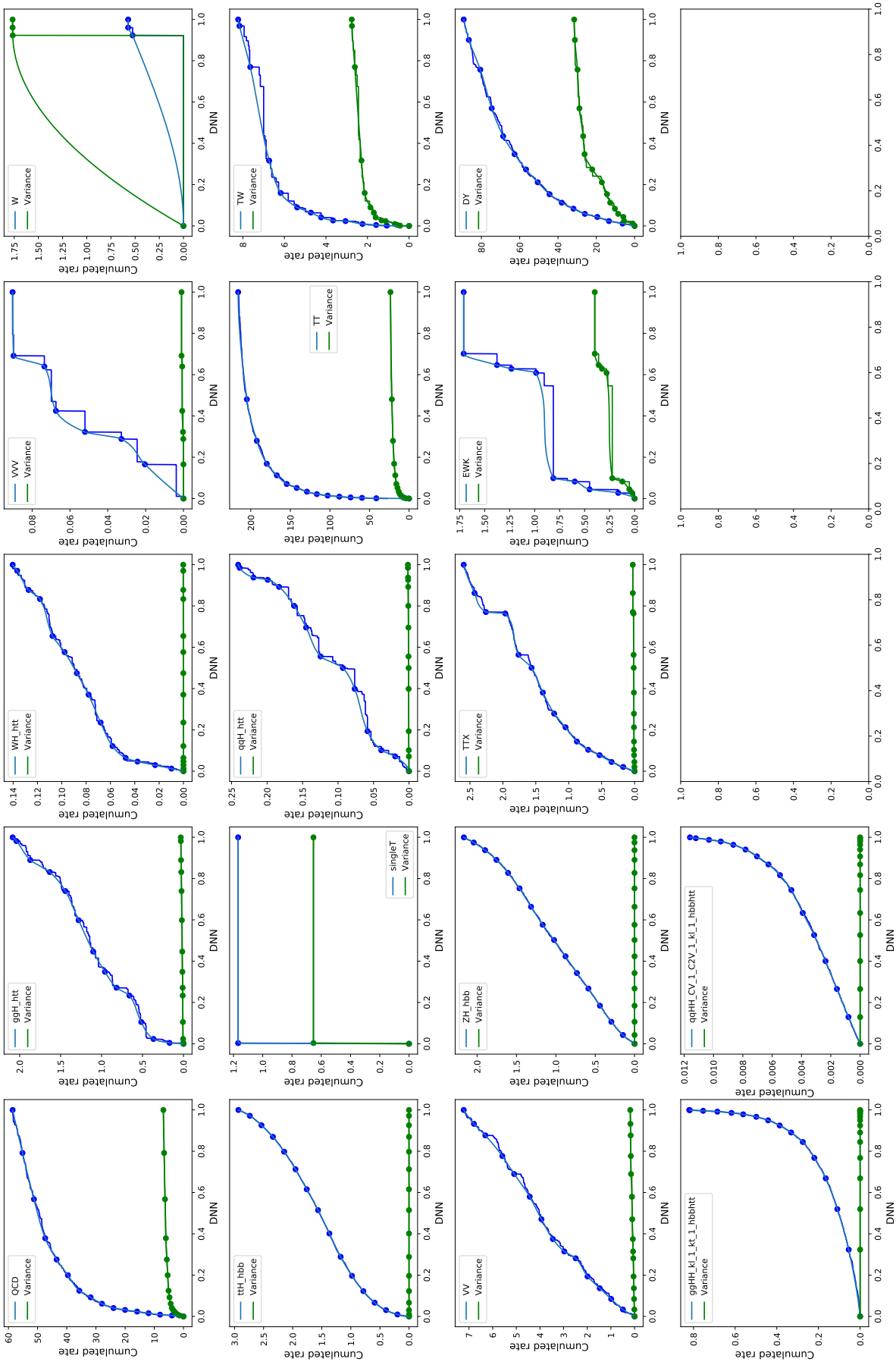


Figure 49: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 15. Shown are for the toy data of toy number 1 the results for all signal and background processes (see legends).

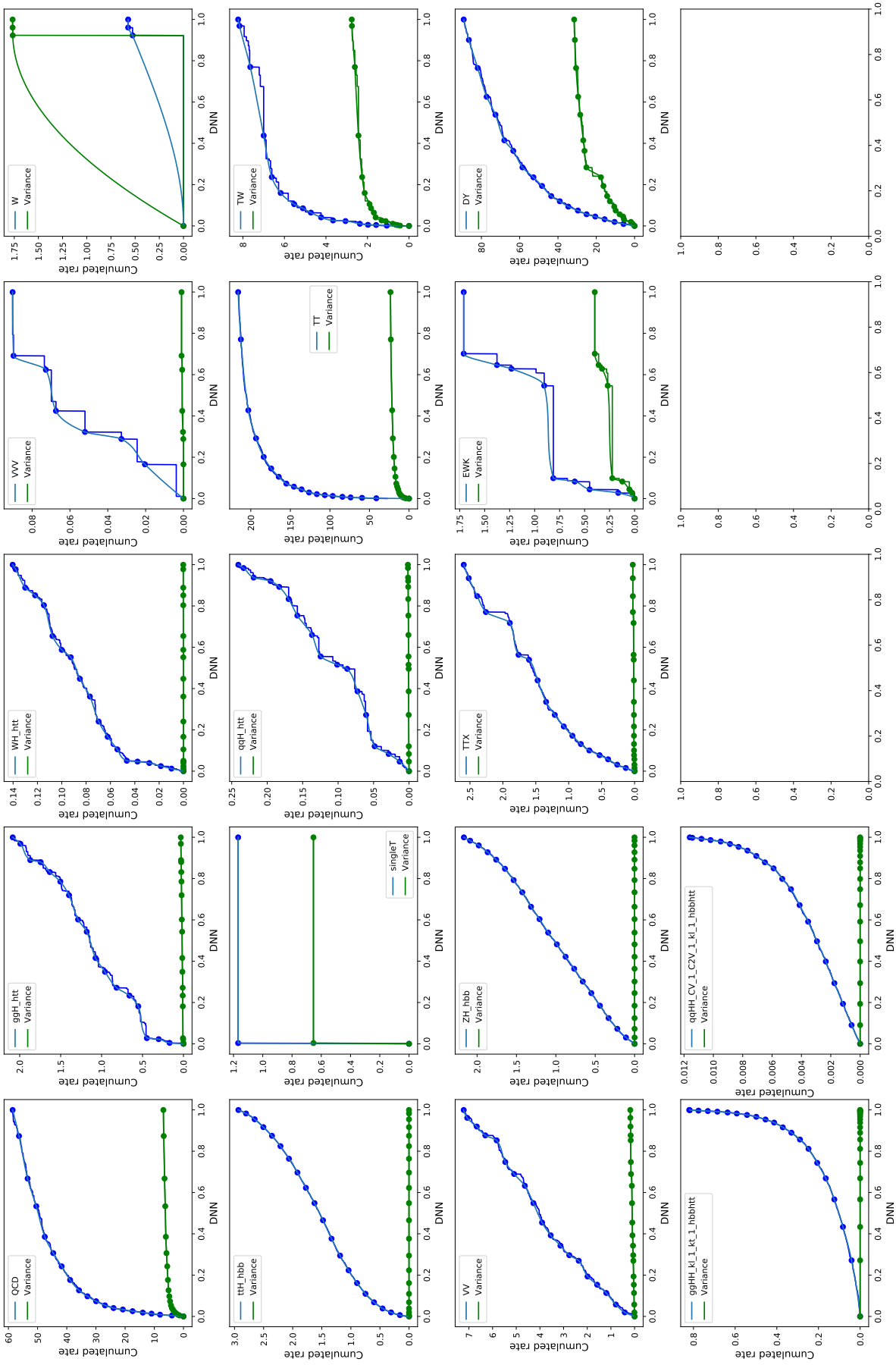


Figure 50: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 20. Shown are for the toy data of toy number 1 the results for all signal and background processes (see legends).

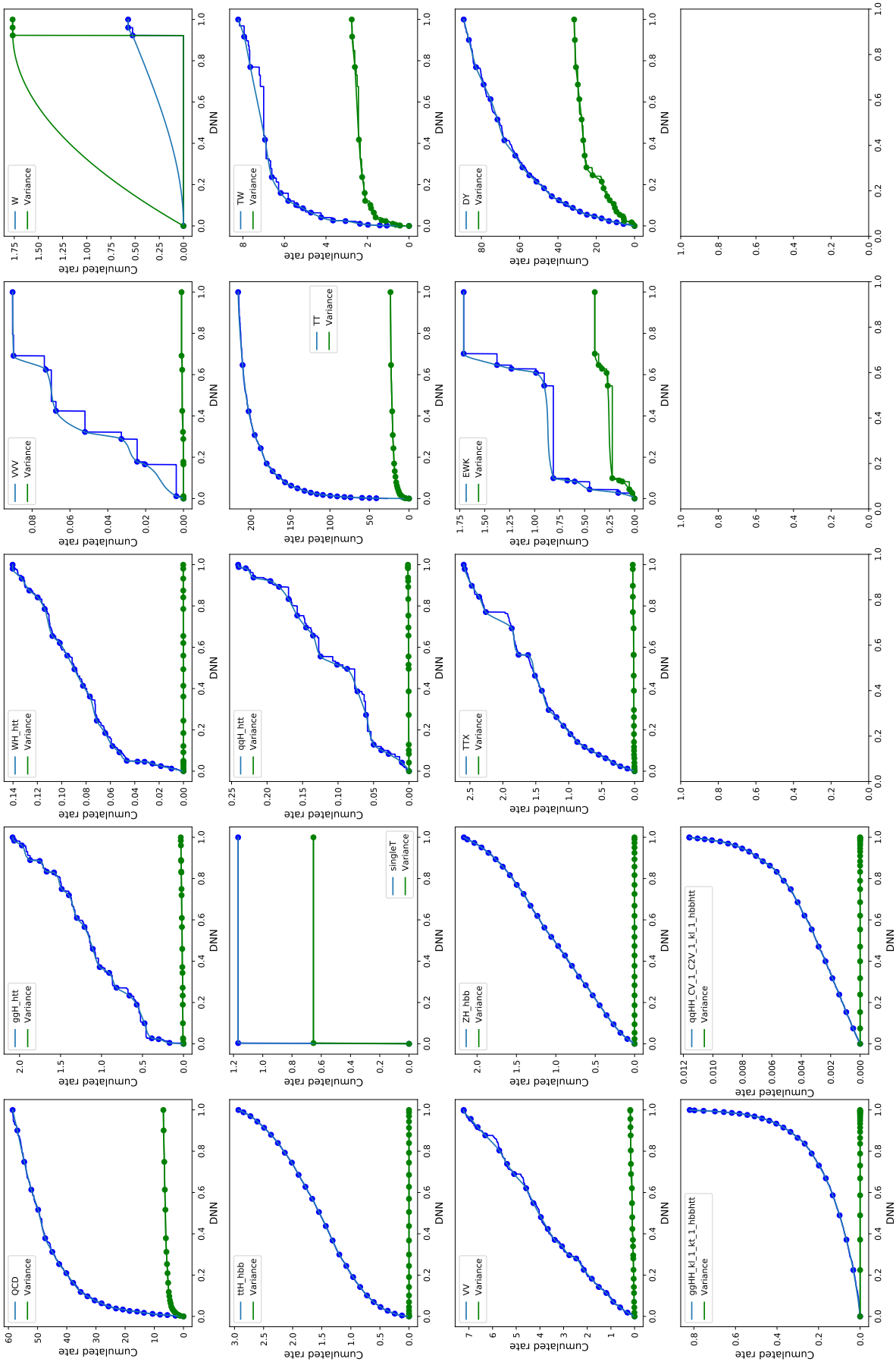


Figure 51: Splines of cumulative distributions of the yields (in blue) and variances (in green) of the discriminator variables for a touch-rate of 25. Shown are for the toy data of toy number 1 the results for all signal and background processes (see legends).

A.3 Distributions of reduced Chi-Squares

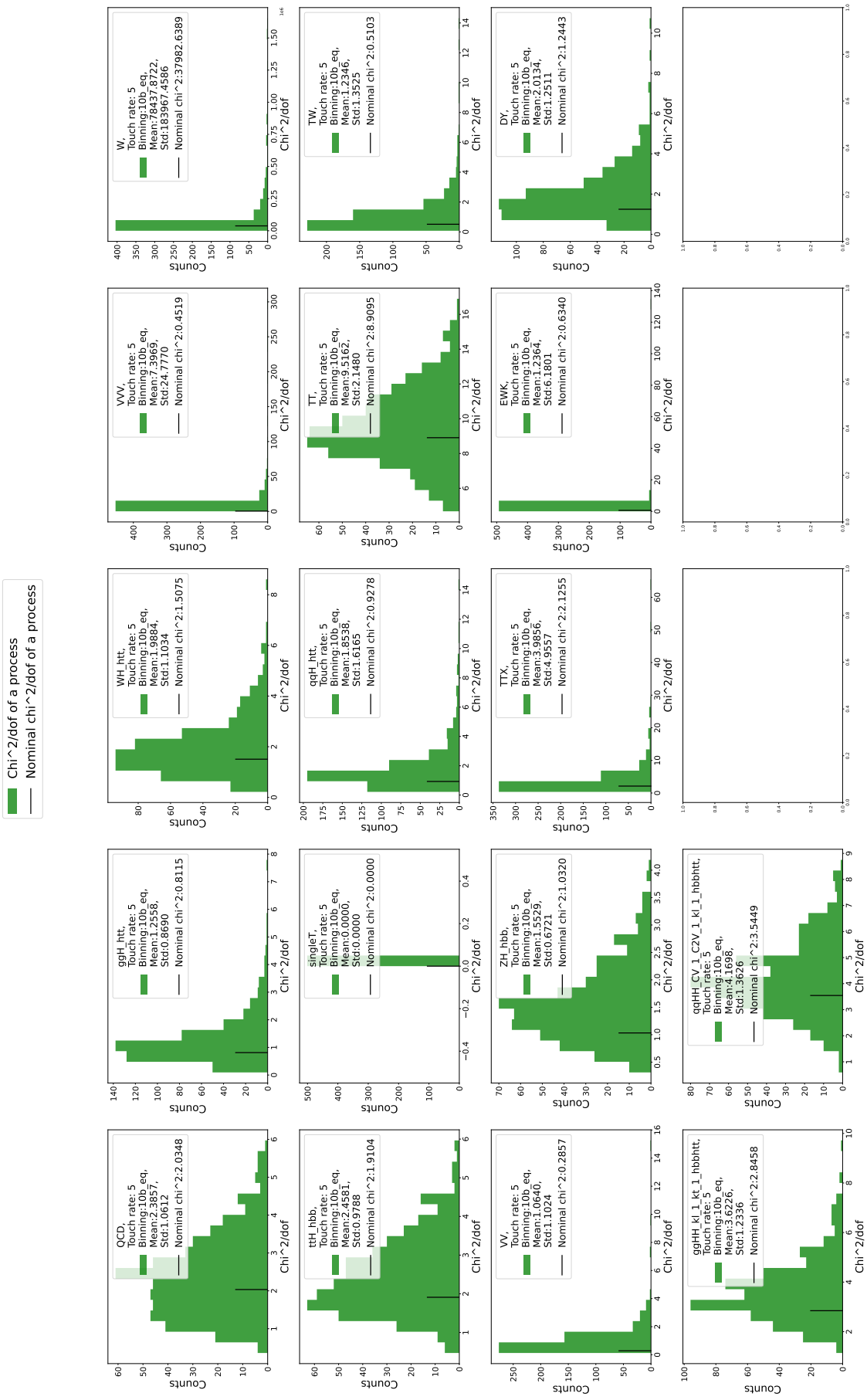


Figure 52: Distributions of reduced chi-squares (Chi^2/dof) for all processes of the splines for a touch-rate of 5 and an equal binning with 10 bins (10b_eq). Shown are toys and original data in green. The original data is highlighted in black.

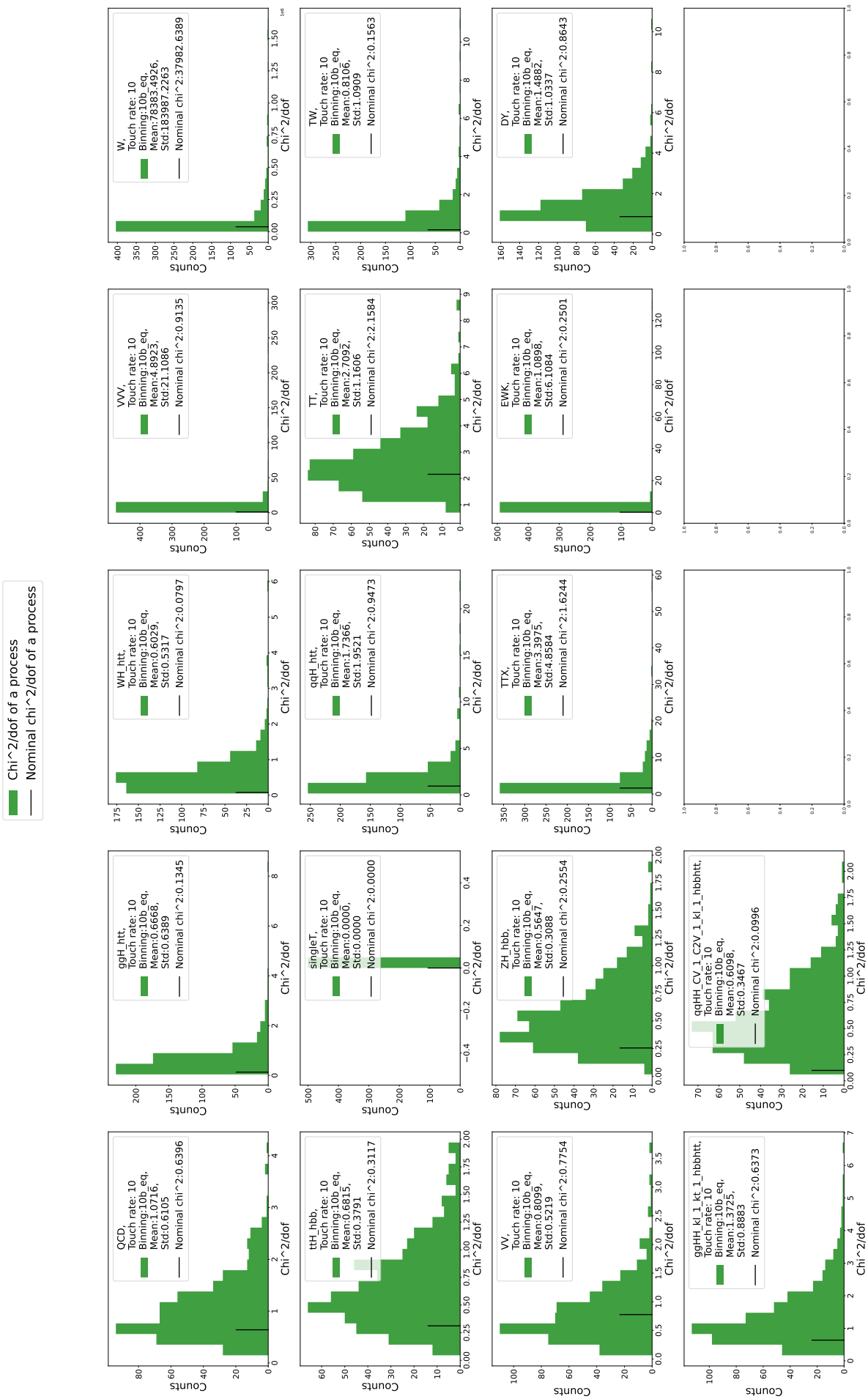


Figure 53: Distributions of reduced chi-squares (Chi^2/dof) for all processes of the splines for a touch-rate of 10 and an equal binning with 10 bins (10b_eq). Shown are toys and original data in green. The original data is highlighted in black.

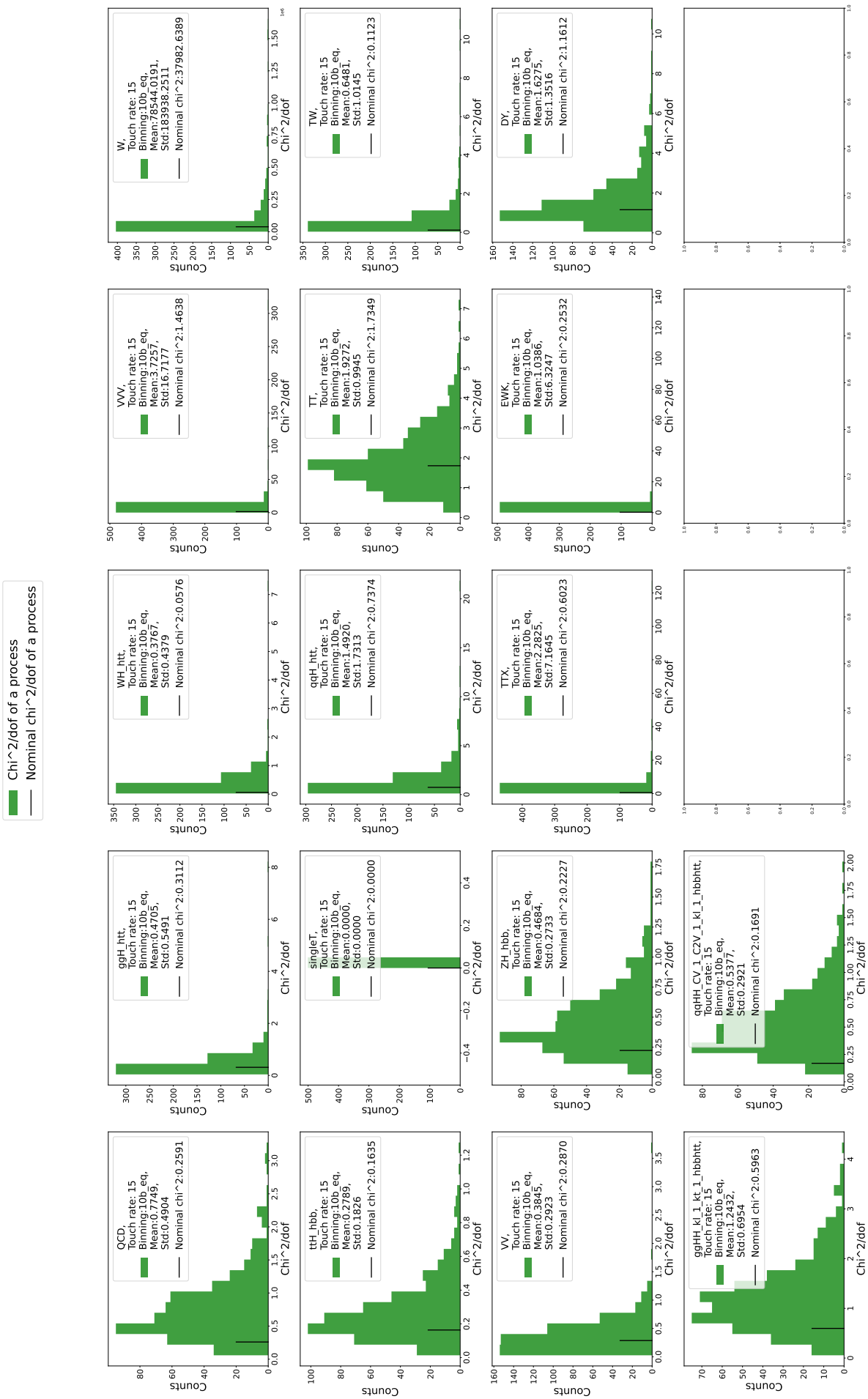


Figure 54: Distributions of reduced chi-squares (Chi^2/dof) for all processes of the splines for a touch-rate of 15 and an equal binning with 10 bins (10b_eq). Shown are toys and original data in green. The original data is highlighted in black.

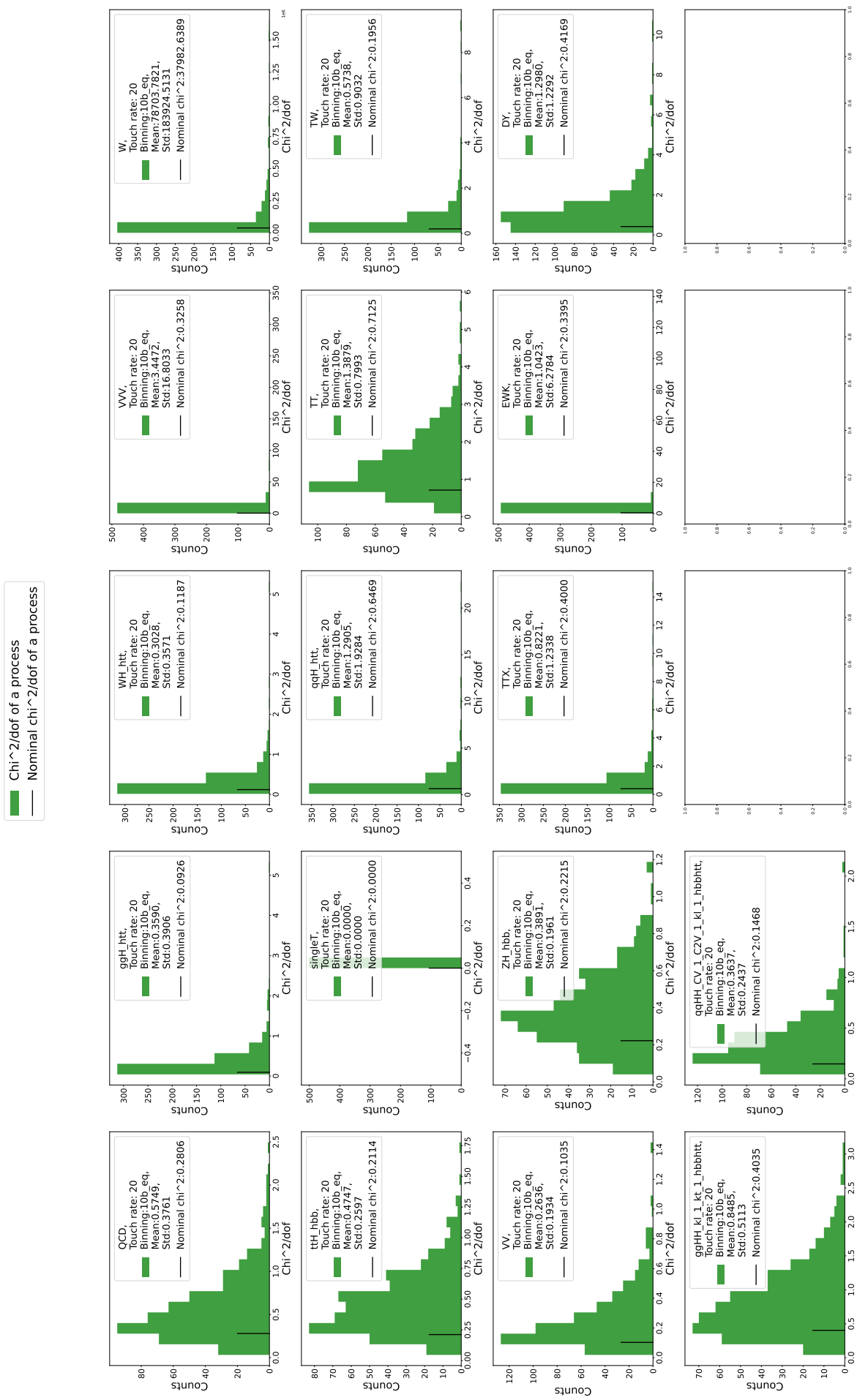


Figure 55: Distributions of reduced chi-squares (Chi^2/dof) for all processes of the splines for a touch-rate of 20 and an equal binning with 10 bins (10b_eq). Shown are toys and original data in green. The original data is highlighted in black.

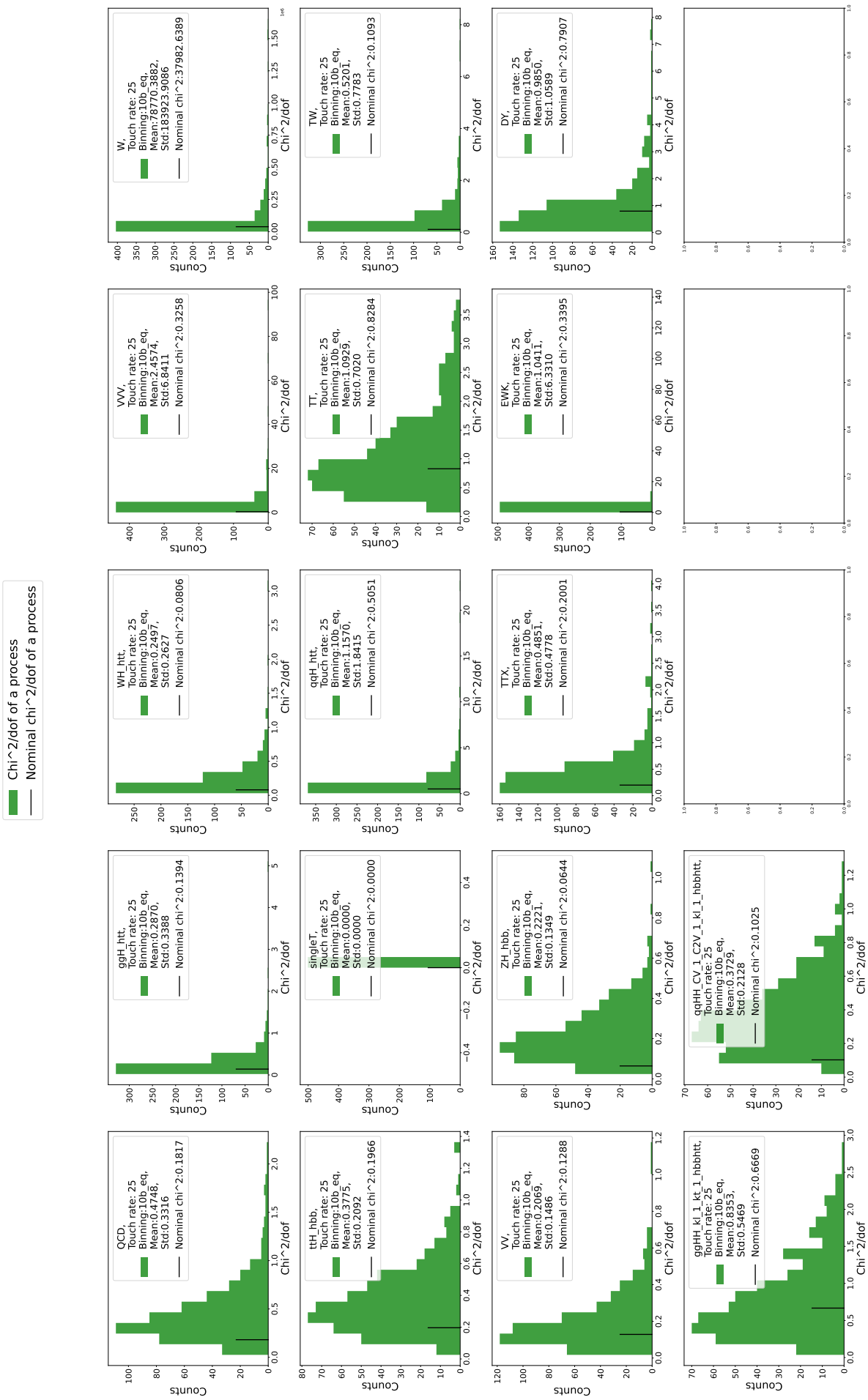


Figure 56: Distributions of reduced chi-squares (Chi^2/dof) for all processes of the splines for a touch-rate of 25 and an equal binning with 10 bins (10b_eq). Shown are toys and original data in green. The original data is highlighted in black.

List of Tables

List of Figures

B References

[1] C. Amendola, K. Androsov, G. Bagliesi, F. Beaudette, R. Bhattacharya, F. Brivio, M. A. Ciocci, O. Davignon, M. R. Di Domenico, V. D Amante, M. Gallinaro, S. Goy Lopez, J. L. Holgado, S. Malvezzi, J. Motta, M. Rieger, R. Salerno, G. Sirois, Y. Strong, and D. Zuolo, *Search for non-resonant higgs boson pair production in bbττ final state with run 2 data*, (2022). HH Approval talk [Online; accessed 26-July-2023].

[2] C. Amendola, K. Androsov, G. Bagliesi, F. Beaudette, R. Bhattacharya, F. Brivio, M. A. Ciocci, O. Davignon, M. R. Di Domenico, V. D'Amante, M. Gallinaro, S. Goy Lopez, J. L. Holgado, S. Malvezzi, J. Motta, M. Rieger, R. Salerno, G. Sirois, Y. Strong, and D. Zuolo, *Search for non-resonant higgs boson pair production in bbττ final state with run 2 data*, (2022). CMS Analysis note CMS AN-18-121.

[3] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: composable transformations of Python+NumPy programs*, 2018.

[4] C. Collaboration, *Search for nonresonant higgs boson pair production in final state with two bottom quarks and two tau leptons in proton-proton collisions at s = 13 tev*, Physics Letters B, 842 (2023), p. 137531. arXiv:2206.09401v2 [hep-ex].

[5] J. S. Conway, *Incorporating nuisance parameters in likelihoods for multisource spectra*, 2011.

[6] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, The European Physical Journal C, 71 (2011).

[7] K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, tech. rep., New York U., New York, 2012.

[8] L. Heinrich, M. Feickert, and G. Stark, *pyhf: v0.6.3*.

[9] F. Marcastel, *CERN's Accelerator Complex. La chaîne des accélérateurs du CERN*, (2013). General Photo.

[10] B. D. Micco, M. Gouzevitch, J. Mazzitelli, and C. Vernieri, *Higgs boson potential at colliders: Status and perspectives*, Reviews in Physics, 5 (2020), p. 100045.

[11] J. B. Mockus and L. J. Mockus, *Bayesian approach to global optimization and application to multiobjective and constrained problems*, Journal of Optimization Theory and Applications, 70 (1991).

[12] J. Neyman, E. S. Pearson, and K. Pearson, *Ix. on the problem of the most efficient tests of statistical hypotheses*, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231 (1933), pp. 289–337.

[13] N. Simpson, *relaxed: version 0.1.3*, Mar. 2022.

[14] A. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, M. Flechl, and et al., *Particle-flow reconstruction and global event description with the cms detector*, Journal of Instrumentation, 12 (2017), p. P10003P10003.

[15] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt,

and SciPy 1.0 Contributors, *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, Nature Methods, 17 (2020), pp. 261–272.

[16] A. Wald, *Tests of statistical hypotheses concerning several parameters when the number of observations is large*, Transactions of the American Mathematical Society, 54 (1943), p. 426.

[17] Wikipedia contributors, *Standard model — Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=Standard_Model&oldid=1157506439, 2023. [Online; accessed 6-June-2023].

[18] S. S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, The Annals of Mathematical Statistics, 9 (1938), p. 60.

Danksagung (Acknowledgements)

Ich bedanke mich bei Prof. Dr. Peter Schleper für das Ermöglichen dieser Arbeit in seiner herausragende Arbeitgruppe, die er mit seinem unermüdlichen Engagement und seiner ansteckenden Freude an Physik leitet.

Bei Dr. Marcel Rieger bedanke ich mich für das interessante Thema der Arbeit und seiner engagierten, geduldigen und sehr kompetenten Betreuung.

Bedanken möchte ih mich auch bei Dr. Philip Keicher, Bogdan Wiederspan, Nathan Prouvost, Tobias Kramer und Dr. Johannes Lange für ihre ständige Hilfsbereitschaft und Kompetenz bei Fragen aller Art.

Bei meinen Bürokollegen und der ganzen Arbeitsgruppe bedanke ich mich für die freundliche und herzliche Arbeitsatmosphäre. Das letzte Jahr hat mir, selbst in den anstrengendsten Phasen, dank euch viel Freude bereitet.
Es gibt viele schöne Zitate aus dem letzten Jahr, aber nur ein relevantes:

> Ich finde das ganze Konzept von Zitaten in Abschlussarbeiten irgendwie mega absurd.
> Dr. Philip Philip

Zum Schluß möchte ich mich noch bei meinen Eltern bedanken, die immer hinter mir standen und mich liebevoll unterstützten. Danke!

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe.

Ich bin mit einer Einstellung in den Bestand der Bibliothek des Fachbereiches einverstanden.

Hamburg, den 05.08.2023 Unterschrift: