IEKP-KA/2015_10

Abschätzung des Untergrundes aus $Z \to \tau\tau$ Ereignissen in $H \to \tau\tau$ Analysen

B.Sc. Benjamin Treiber

Masterarbeit

an der Fakultät für Physik
des Karlsruher Instituts für Technologie (KIT)

Referent: Prof. Dr. Günter Quast
Institut für Experimentelle Kernphysik

Korreferent: Prof. Dr. Ulrich Husemann
Institut für Experimentelle Kernphysik

31. Juli 2015

Zusammenfassung

Die Teilchenphysik wurde in den vergangenen Jahrzehnten von der Suche nach den vom Standardmodell der Elementarteilchenphysik postulierten Teilchen angetrieben. Das Standardmodell bildet die Grundlage des heutigen Verständnisses der Teilchenphysik. Es bietet eine Umfassende Beschreibung fundamentaler Prozesse in unserem Universum und wurde in zahlreichen Experimenten bestätigt.

Das Higgs Boson wird im Standardmodell benötigt, um die Existenz der massiven Eichbosonen der schwachen Wechselwirkung mit dem Erhalt lokaler Eichsymmetrien zu vereinbaren. Die Suche nach diesem Teilchen fand ihren Höhepunkt im Juli 2012, als am europäischen Kernforschungszentrum CERN die Entdeckung eines neuen Bosons mit Eigenschaften des Standardmodell Higgs Bosons bekannt gegeben wurde.

Die Erfolgsgeschichte des Standardmodells ist eng verknüpft mit der Erfolgsgeschichte vieler Beschleuniger- und Detektorexperimente mit Forschungsbeiträgen aus der ganzen Welt. Diese Experimente ermöglichten es, Teilchen mit immer größeren Energien zur Kollision zu bringen und dadurch neue, schwerere Teilchen zu erzeugen und zu studieren. Dies ermöglichte erst die Validierung des Standardmodells sowie die Entdeckung der postulierten Teilchen.

Im Zuge größerer und komplexerer Beschleuniger- und Detektorexperimente trugen auch neue Methoden der Datenanalyse zu den erfolgreichen Entdeckungen bei. Ein Ziel der Datenanalyse ist es, die Ausbeute eines Datensatzes zu maximieren, indem man eine geeignete Kombination von Parametern findet, um zwischen Signal- und Untergrundereignissen zu unterscheiden. Ein Beispiel dafür sind Multivariate Analysemethoden. Diese Methoden erlauben durch die Kombination verschiedener Parameter und deren Korrelationen Hyperflächen zu finden, welche eine gute Separation zwischen Signal und Untergrund erlauben.

Voraussetzung für diese Analysemethoden ist jedoch, dass ein solcher Satz an Parametern, der eine Unterscheidung zwischen Signal und Untergrund ermöglicht, überhaupt existiert. Zum Beispiel ist der Zerfall eines Standardmodell Higgs Bosons in zwei Tauonen ($H \to \tau\tau$) kinematisch und topologisch fast identisch zum Zerfall des Z Bosons in zwei Tauonen ($\mathrm{Z} \to \tau\tau$). In diesem Fall findet sich kein Satz an Parametern, der eine effektive Unterscheidung zwischen dem $H \to \tau\tau$ Signal und dem $\mathrm{Z} \to \tau\tau$ Untergrund ermöglicht. In einem solchen Fall müssen andere Methoden zur Unterdrückung von Untergrundereignissen gefunden werden. Ein Beispiel für eine solche Methode ist das sogenannte *Embedding*, das in Analysen des Compact Muon Solenoid Detektor Experimentes (CMS) angewendet wird. Diese Methode erlaubt es, systematische Unsicherheiten im $\mathrm{Z} \to \tau\tau$ Untergrund zu reduzieren und dadurch indirekt die Signifikanz von Signalen zu erhöhen.

Da $Z \to \tau\tau$ Zerfälle, unter anderem wegen der hadronischen Zerfälle der Tauonen, schwierig zu selektieren sind, ist es auch schwierig den $Z \to \tau\tau$ Untergrund auf Basis aufgezeichneter Daten abzuschätzen. Eine rein auf Monte Carlo (MC) Simulation von $Z \to \tau\tau$ Ereignissen basierende Untergrundabschätzung unterliegt jedoch zusätzlichen systematischen Unsicherheiten, zum Beispiel in der Energiekalibration von Jets und durch Ungenauigkeiten in der Detektorsimulation und -auslese.

Das Embedding macht sich die Vorteile von beiden Methoden der Untergrundabschätzung zu Nutze. Die Methode ist in Abbildung 1 schematisch dargestellt und lässt sich in den folgenden Schritten zusammenfassen:

1. **Selektion von $Z \to \mu\mu$ Zerfällen**
   Ereignisse mit zwei Myonen werden aus aufgezeichneten Daten selektiert. Beide Myonen müssen dabei gewisse Mindestanforderungen unter anderem bezüglich der Anzahl von Datenpunkten im Spurdetektor und den Myonenkammern erfüllen. Dadurch wird sichergestellt, dass der gemessene Impuls präzise gemessen wurde. Zusätzlich müssen die rekonstruierten Myonen gut isoliert sein. Dadurch werden Fehlidentifikationen zum Beispiel von geladenen Pionen, die als Myon rekonstruiert wurden, unterdrückt. Diese ausgewählten $Z \to \mu\mu$ Ereignisse bilden die Basis für das Embedding und werden im Folgenden als *ursprüngliches* Ereignis bezeichnet.

2. **Entfernung der Myonen aus dem ursprünglichen Ereignis**
   Die rekonstruierten Myonen sowie deren Spuren und Kalorimetertreffer werden aus dem ursprünglichen Ereignis entfernt.

3. **Simulation eines separaten $Z \to \tau\tau$ Ereignisses**
   Ein separates Ereignis mit zwei simulierten Tauonen wird generiert. Die Tauonen erhalten zunächst den selben Vierimpuls wie die zuvor entfernten Myonen. Die größere Masse der Tauonen wird dann durch eine Korrektur auf den Vierimpuls berücksichtigt, sodass die invariante Masse $m_{\tau\tau}$ des Ereignisses unverändert bleibt. Da Myonen im $Z \to \mu\mu$ Ereignis bereits Final State Radiation emittieren, ist diese im simulierten $Z \to \tau\tau$ Ereignis deaktiviert, um doppelte Emissionen und Verzerrungen der Kinematik zu vermeiden. Das simulierte Ereignis durchläuft dann die Detektorsimulation und die Ereignisrekonstruktionsalgorithmen. Dieses separate Ereignis wird im Folgenden als *simuliertes* Ereignis bezeichnet.

4. **Einbettung in das ursprüngliche Ereignis**
   Das Signal aus dem simulierten Ereignis wird dann in den nach der Entfernung des Myonensignals verbleibenden Rest des ursprünglichen Ereignisses eingebettet. Dadurch entsteht ein Daten/Monte Carlo Hybrid Ereignis, ein *eingebettetes* oder auch *embedded* Ereignis.

Die Embedding Methode macht sich die Vorteile von datenbasierter und Monte Carlo basierter Untergrundabschätzung zunutze. Die $Z \to \mu\mu$ Ereignisse können

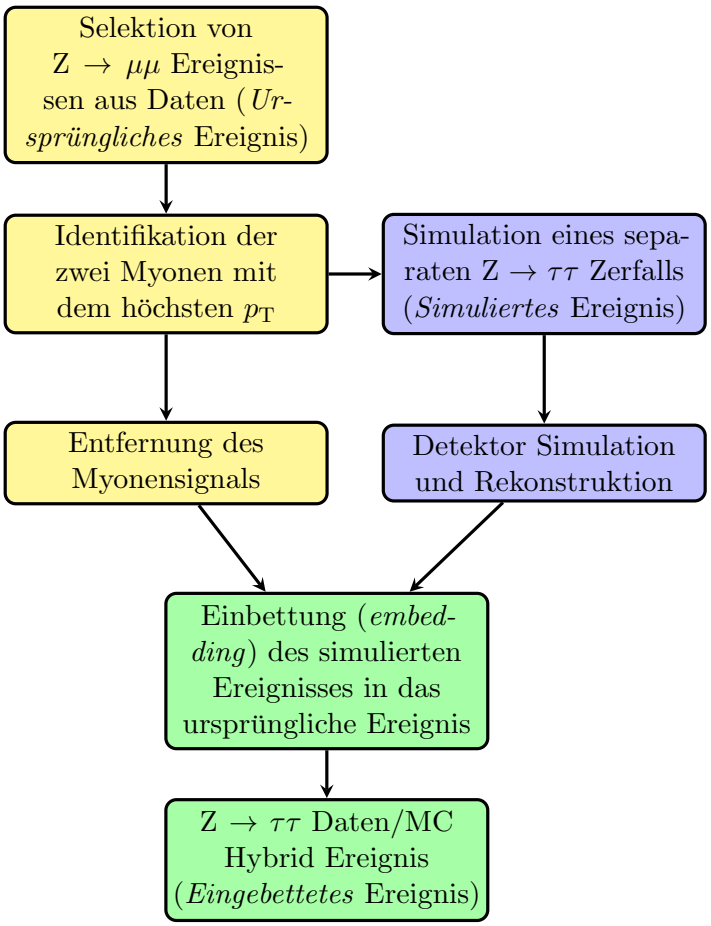


Abbildung 1: Schematische Darstellung der Embedding Methode. Das *ursprüngliche* aufgezeichnete $Z \to \mu\mu$ Ereignis ist in gelb dargestellt, das neu erstellte *simulierte* Ereignis in blau und das kombinierte, *eingebettete* Ereignis in grün. Ausgehend von selektierten $Z \to \mu\mu$ Ereignissen werden die zwei Myonen mit größtem Transversalimpuls p_T identifiziert. Die rekonstruierten Myonen sowie deren Signal in Kalorimeter und Spurdetektor werden aus dem ursprünglichen Ereignis entfernt. Parallel dazu wird ein neues $Z \to \tau\tau$ Monte Carlo Ereignis erzeugt. Die darin simulierten Tauonen erhalten die gleiche Position und Richtung des Viererimpulses wie die zuvor rekonstruierten Myonen. Lediglich der Betrag des Impulses wird leicht korrigiert, um den unterschiedlichen Massen von Myon und Tauon zu berücksichtigen. Das simulierte Ereignis durchläuft dann die Detektorsimulation und die Ereignisrekonstruktion. Anschließend wird das simulierte Ereignis in den Rest des ursprünglichen Ereignisses eingebettet. Das Resultat ist ein sogenanntes *eingebettetes* Ereignis, ein Daten/MC Hybrid Ereignis.

mit großer Reinheit und Effizienz selektiert werden. Dies führt auch dazu, dass die $Z \to \tau\tau$ Zerfälle im eingebetteten Ereignis genauer beschrieben sind, als wenn man sie direkt aus Daten selektiert hätte. Gleichzeitig sind Unsicherheiten, wie man sie aus einer reinen Monte Carlo Simulation erwarten würde, stark unterdrückt, da der größte Teil des Ereignisses nach der Einbettung noch immer aus Daten stammt.

In der Entwicklung eines Embedding Algorithmus muss zunächst eine Stufe der Ereignisrekonstruktion gefunden werden, in der die Einbettung von simuliertem Ereignis in das ursprüngliche Ereignis stattfinden kann. Das theoretisch niedrigste Niveau hierfür wäre auf Ebene der digitalisierten Detektorrohdaten, da dies auch die unterste Ebene der Simulation ist. Bedingt durch Einschränkungen bezüglich der verfügbaren Speicherkapazität und -bandbreite können die Ereignisse jedoch nicht auf dieser Ebene aufgezeichnet werden.

Die niedrigste Ebene, auf der die Kollisionsdaten bereitgestellt werden, ist die Ebene von rekonstruierten Spurdetektor- und Kalorimetertreffern. Das simulierte Ereignis wird bis zu diesem Schritt rekonstruiert und dann mit dem ursprünglichen Ereignis kombiniert. Diese Embedding Methode wird *Rec-Hit* Embedding (RH Embedding) genannt.

Eine zweite Ebene, auf der die Einbettung durchgeführt werden kann, ist die Ebene rekonstruierter Teilchen. In CMS werden Teilchen mit einem sogenannten Teilchenfluss Algorithmus (*particle flow* Algorithmus) rekonstruiert. Entsprechend werden das ursprüngliche und das simulierte Ereignis auf Ebene der rekonstruierten Teilchen zusammengeführt. Diese Methode wird *Particle Flow* Embedding (PF Embedding) genannt.

Die Anforderung an die Isolation in der Selektion der ursprünglichen $Z \to \mu\mu$ Ereignisse ist, dass die Summe des Impulses aus Aktivität geladener hadronischer Teilchen in einem Kegel um die Flugrichtung eines jeweiligen Myons geringer ist, als 10 % des gemessenen Impulses des Myons. Ist dieses Kriterium nicht erfüllt, wird das rekonstruierte Myon verworfen. Dieses Kriterium führt dazu, dass die selektierten Ereignisse überdurchschnittlich gut isoliert sind. Dies wirkt sich auch auf das Embedding aus, da das Signal aus den simulierten Ereignissen in überdurchschnittlich gut isolierte Umgebungen des Detektors eingebettet werden. Dies führt zu erhöhten Selektionseffizienzen.

Um diese systematische Abweichung zu reduzieren, wurde eine Spiegeltransformation angewendet. Diese spiegelt den Transversalimpulsvektor der Myonen an der von Z Boson Impuls und dem Impuls des ursprünglich kollidierenden Protons aufgespannten Ebene. Der Z Boson Zerfall bleibt durch diese Transformation kinematisch unverändert. Im simulierten Ereignis werden die Teilchen dann mit dem gespiegelten Myonenimpuls eingebettet und befinden sich daher in einem anderen Bereich des Detektors, der weniger stark von der $Z \to \mu\mu$ Ereignisselektion beeinflusst wurde.

Im Rahmen der Vorbereitungen für die zweite Datennahmeperiode am *Large Hadron Collider* (LHC), wurden Rekonstruktionsalgorithmen im Softwarepaket CMSSW der CMS Kollaboration verändert und angepasst. Dies machte auch Anpassungen und eine erneute Untersuchung der Embedding Algorithmen notwendig.

Ergebnisse der Einbettung von Myonen

Um zu überprüfen, ob die Änderungen erfolgreich umgesetzt wurden, wurde das Embedding modifiziert, sodass Myonen anstatt Tauonen im simulierten Ereignis generiert und nach Rekonstruktion eingebettet werden. Durch die Ersetzung der Myonen aus dem ursprünglichen Ereignis mit neu simulierten Myonen lassen sich systematische Abweichungen und Verzerrungen aufgrund der Embedding Algorithmen selbst am besten untersuchen.

Da zu Beginn dieser Arbeit die zweiten Datennahmeperiode noch nicht begonnen hatte, wurden alle Untersuchungen auf Basis von MC simulierten $Z \to \mu\mu$ und $Z \to \tau\tau$ Datensätzen durchgeführt.

Es wurden vier verschiedene Embedding Methoden verglichen. Das PF und RH Embedding wurde jeweils mit und ohne Spiegelung der eingebetteten Teilchen durchgeführt. Da im Embedding Abweichungen in der Selektionseffizienz in Abhängigkeit der Anzahl an zeitgleich ablaufenden Proton-Proton Kollisionen (nPU) erwartet werden, wurde die Selektionseffizienz der Embedding Algorithmen in Abhängigkeit dieser Anzahl untersucht. Auftretende Abweichungen wurden anhand des Vergleiches des Selektionseffizienz mit einem $Z \to \mu\mu$ Datensatz quantifiziert.

Die Ergebnisse der Einbettung von Myonen sind in Abbildung 2a dargestellt.

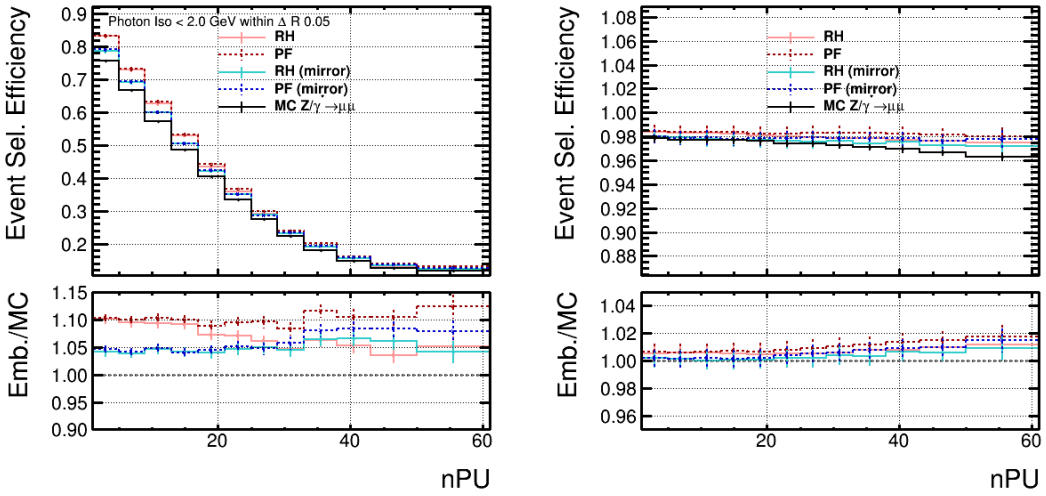


Abbildung 2: Die Selektionseffizienz der vier untersuchten Myon Embedding Methoden im Vergleich zu einem $Z \to \mu\mu$ Datensatz zur Validierung ist in Abbildung 2a dargestellt. Die beste Übereinstimmung mit dem Validierungsdatensatz zeigt das RH Embedding mit Spiegelung der eingebetteten Teilchen. Hier liegen die Abweichungen in Abhängigkeit von nPU zwischen 5 % und 7 %. Abbildung 2b zeigt die Abweichungen der Selektionseffizienz ohne Anwendung der $\Delta\beta$-korrigierten Isolation. Ohne dieses Kriterium sind die Abweichungen in allen Embedding Methoden geringer als 2 %.

Die Abweichungen lassen sich hauptsächlich auf die $\Delta\beta$-korrigierte Isolation der Myonen zurückführen, die sich aus den Transversalimpulsen geladener Hadronen, neutraler Hadronen und Photonen zusammensetzt. Dies geht auch aus Abbildung 2b hervor. Diese zeigt die Abweichungen der Selektionseffizienz der vier Embedding Algorithmen wenn alle Kriterien der Basisselektion außer das Isolationskriterium angewendet werden. Die Abweichungen ohne die Berücksichtigung der Isolation sind kleiner als 2 %.

Wie zuvor beschrieben, sind die in den ursprünglichen Ereignissen enthaltenen Myonen überdurchschnittlich gut isoliert. Dies führt zu einer 10 % höheren Selektionseffizienz des PF und RH Embeddings ohne die Spiegelung. In den gespiegelten Embedding Methoden wird diese Differenz auf 2 % reduziert.

In der Isolationskomponente aus neutralen Hadronen kommt es zu einer systematischen Abweichung in gespiegelten PF Embedding und im ungespiegelten RH Embedding. Diese Abweichung hat ihren Ursprung im Teilchenrekonstruktionsalgorithmus particle flow. Dieser modifiziert die Kollektion der rekonstruierten neutralen Hadronen im Ereignis in Anwesenheit eines Myons. Die Modifikation wird im RH Embedding ein zweites Mal bei der Rekonstruktion der Teilchen im eingebetteten Ereignis angewendet. Im PF Embedding erfolgt diese Teilchenrekonstruktion im simulierten Ereignis und beeinflusst daher keine neutralen Hadronen im ursprünglichen Ereignis. Im ungespiegelten RH Embedding werden diese Modifikationen doppelt an gleicher Stelle angewendet, im gespiegelten PF Embedding hingegen gehen die Änderungen durch die Spiegelung verloren. Im RH Embedding führt dies zu einer mit zunehmender Anzahl paralleler Proton-Proton Kollisionen im Ereignis zu 6 % niedrigeren Selektionseffizienz als im Validierungsdatensatz bei ungefähr 50 parallelen Kollisionen. Dieser Abwärtstrend ist für das RH Embedding ohne Spiegelung auch in Abbildung 2a erkennbar. Im gespiegelten PF Embedding führt dies zu bis zu 2 % höheren Selektionseffizienzen bei ungefähr 50 parallelen Kollisionen. Dieser leichte Aufwärtstrend ist ebenfalls in Abbildung 2a zu sehen.

Abweichungen aufgrund der Isolationskomponente von Photonen hängen im Wesentlichen mit der Emission von Final State Radiation der Myonen zusammen. Insbesondere mit der Spiegelung der eingebetteten Teilchen gehen diese teils hochenergetischen Photonen verloren, da sie nicht den Myonen zugeordnet und folglich auch nicht mit gespiegelt werden können. Aufgrund der Korrelation zwischen der deponierten Photonenenergie nahe des Myons und dem Auftreten von Final State Radiation lässt sich der Einfluss der Final State Radiation etwas reduzieren. Hierfür wurde gefordert, dass nicht mehr als 2 GeV Transversalimpuls von Photonen innerhalb eines η-ϕ Kegels kleiner $\Delta R = 0.05$ um die Flugrichtung des Myons herum gemessen wurde. Die Herausnahme von Ereignissen, bei denen dieses Kriterium nicht erfüllt war, führte zu einer deutlichen Reduktion der durch Final State Radiation bedingten Abweichung zwischen gespiegelten und ungespiegeltem Embedding.

Ergebnisse der Einbettung von Tauonen

Das Embedding von Tauonen wurden im $\mu\tau_\mathrm{h}$ Endzustand der zwei Tauonen im Ereignis untersucht. In diesem Endzustand zerfällt ein Tauon leptonisch in ein Myon, das zweite Tauon hadronisch. Die Selektionseffizienzen der Embedding Methoden in diesem Zerfallskanal im Vergleich zu einem $\mathrm{Z} \to \tau\tau \to \mu\tau_\mathrm{h}$ Validierungsdatensatz sind in Abbildung 3a dargestellt.

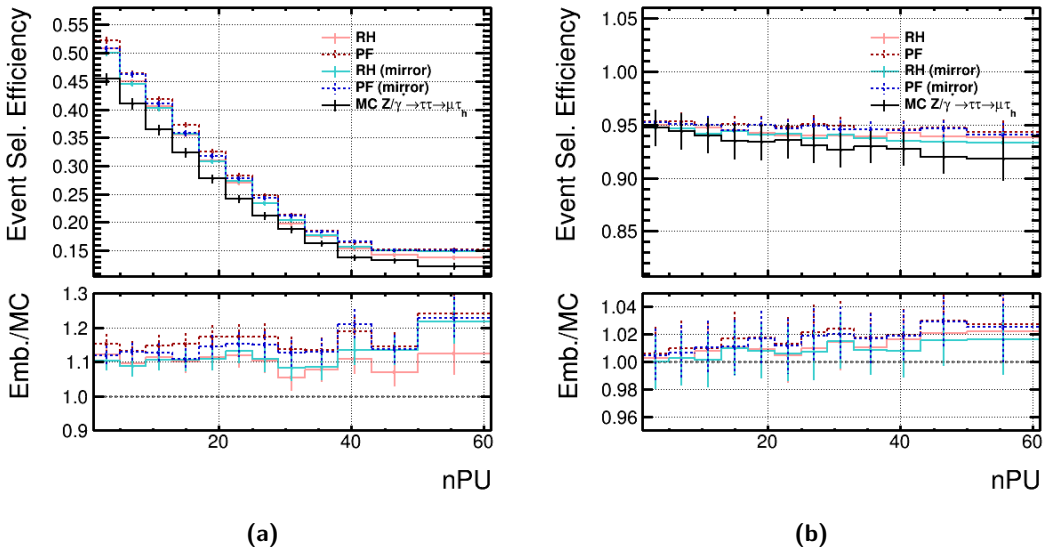


Abbildung 3: Die Selektionseffizienz der vier untersuchten Tau Embedding Methoden im Vergleich zu einem $\mathrm{Z} \to \tau\tau \to \mu\tau_\mathrm{h}$ Datensatz zur Validierung ist in Abbildung 3a dargestellt. Alle Embedding Algorithmen zeigen eine um 10 % bis 20 % höhere Selektionseffizienz als der Vergleichsdatensatz. Abbildung 3b zeigt die Abweichungen der Selektionseffizienz ohne Anwendung von Isolationskriterien für die rekonstruierten Myonen und hadronisch zerfallenen Tauonen. Die Abweichungen in allen Embedding Methoden sind in diesem Fall geringer als 3 %.

Die vier verschiedenen Embedding Methoden zeigen eine zwischen 10 % und 20 % höhere Selektionseffizienz als der Kontrolldatensatz. Die Abweichungen aufgrund des rekonstruierten Myons darin beträgt bis zu 7 %. Im Tau Embedding reduziert die Forderung, dass sich nicht mehr als 2 GeV Transversalimpuls von Photonen innerhalb eines η-ϕ Kegels kleiner $\Delta R = 0.05$ um die Myonen befindet, die Abweichungen der gespiegelten Embedding Methoden aufgrund von Final State Radiation um bis zu 3 %. Die Abweichungen aufgrund der Isolationskomponente der hadronisch zerfallenden Tauonen τ_h liegen zwischen 10 % und 15 %.

Auch hier gehen Abweichungen hauptsächlich auf Unterschiede in der Isolationskomponente zurück. Abbildung 3b zeigt die Selektionseffizienz des Tau Embeddings ohne die Anwendung von Isolationskriterien. Die Abweichungen zwischen Embedding und Kontrolldatensatz liegen unter 3 %.

Ausblick

Die Ursache für Abweichungen der Selektionseffizienzen, insbesondere im Embedding von Myonen, sind gut verstanden. Da die größten Abweichungen mit der Isolation der eingebetteten Teilchen zusammenhängt, sollte dieser Aspekt des Embeddings weiterhin studiert und genauer untersucht werden.

Die Isolationskomponente aus geladenen Hadronen zeigt auch mit Anwendung der Spiegeltransformation noch Abweichungen hin zu besserer Isolation. Grund hierfür könnte ein auch mit der Spiegelung bestehender Einfluss der Selektion der ursprünglichen $Z \to \mu\mu$ Ereignisse sein. Nach der Spiegelung kann das für ein Myon eingebettete Teilchen in die ursprüngliche Richtung des zweiten Myons im Ereignis zeigen. Dadurch würde dieses eingebettete Teilchen sich wiederum in einem überdurchschnittlich gut isolierten Bereich des Detektors befinden. Ob dies die Ursache für die fortbestehende Abweichung ist, könnte überprüft werden, indem Ereignisse bei denen sich die für die Isolationsberechnung verwendeten η-ϕ Kegel von gespiegelten eingebetteten Teilchen mit den entsprechenden Isolationskegeln der ursprünglichen Myonen im Ereignis überschneiden aus den Datensätzen herausgenommen und die Isolationskomponente der geladenen Hadronen erneut verglichen werden.

Aus der Isolationskomponente neutraler Hadronen hervorgehende Abweichungen werden voraussichtlich ab der Version CMSSW 7.6 behoben sein, da der Prozess der die Abweichungen verursacht ab dieser Version von CMSSW entfernt und anderweitig berücksichtigt wird.

Weiterhin sollten die Effekte auf Grund von Final State Radiation der Myonen genauer studiert werden. Hierfür bietet sich an, die in MC Simulationen verfügbaren Informationen der Ereignisgeneratoren auszunutzen und Ereignisse in denen die Myonen Final State Radiation emittiert haben herauszufiltern und die Embedding Methoden in einem Datensatz frei von Final State Radiation anzuwenden. So ließe sich im Vergleich zu den Methoden mit dieser Strahlung Störquellen und die Größe der Abweichungen die auf Final State Radiation beruhen abschätzen.

Die Ursache der Abweichungen in der Isolation der τ_h muss weitergehend studiert werden. Die Abweichungen sind sehr ähnlich für alle Embedding Methoden, unabhängig von der Spiegelung. Daher ist die Ursache hierfür wahrscheinlich in der Simulation des Zerfalls der Tauonen zu finden.

IEKP-KA/2015_10

Estimation of the Background from $Z \to \tau\tau$ in $H \to \tau\tau$ Analyses

B.Sc. Benjamin Treiber

Masterthesis

at the Department of Physics
of the Karlsruhe Institute of Technology (KIT)

Reviewer: Prof. Dr. Günter Quast
Institute of Experimental Nuclear Physics

Second Reviewer: Prof. Dr. Ulrich Husemann
Institute of Experimental Nuclear Physics

31. Juli 2015

Contents

1 Introduction

In the past decades, the field of particle physics was driven by the search for the leptons, gauge bosons and quarks postulated by the Standard Model of particle physics. The Standard Model builds the foundation of our present understanding of particle physics. It provides a comprehensive description of fundamental processes and forces in our Universe and has been validated in numerous experiments.

The Higgs boson is required in the Standard Model to unify the existence of massive gauge bosons with the preservation of local gauge symmetries. The search for it culminated in July 2012, when the first evidence of the existence of a boson with properties matching the ones of the Standard Model Higgs boson was announced at CERN [1, 2].

The story of the success of the Standard Model is closely related to the success of many accelerator and detector experiments with contributions from all around the world. The accelerators, like the Large Hadron Collider, collided particles at increasing centre-of-mass energies $\sqrt{s}$. This made it possible to create new particles of higher masses. During extended data taking periods, large amounts of collision data could be recorded and the predictions of the Standard Model tested with increasing precision.

In the shadow of the accelerators and detectors, also new data analysis techniques contributed to the discoveries. Their goal is to maximally exploit a dataset by finding the best selection of parameters that allows to differ between signal and background contributions. One example for this are multivariate analysis strategies. These exploit the discriminating power of kinematic properties and event topologies and are effective methods to take correlations between parameters into account. This way, the multivariate analysis methods are able to reliably find hyperplanes with a good discriminating power in high dimensional spaces that are otherwise only difficult to grasp.

A prerequisite of the multivariate analysis methods is the existence of a set of discriminating variables. In some cases, no such set of parameters can be found. For example the decay of a Higgs boson into two τ-leptons is kinematically and topologically almost identical to the decay of a Z boson into two τ-leptons. Therefore, new methods need to be studied to decrease the uncertainties on the $\mathrm{Z} \to \tau\tau$ background and thereby reduce its impact on analyses. One method that is able to do this is the so called *embedding* procedure studied in this thesis.

The embedding procedure combines the advantages of data driven and Monte Carlo simulation based background estimations. It uses recorded $\mathrm{Z} \to \mu\mu$ events that can be selected with very high efficiency and purity. The muons from these events are then removed and replaced by simulated τ-leptons. This way, a more precise

description of $Z \to \tau\tau$ decays is achieved compared to a purely data or Monte Carlo driven background estimation method.

In preparation for the second data taking period of the Large Hadron Collider, various adjustments were introduced into the software framework of the Compact Muon Solenoid detector experiment. This made several adjustments in the embedding algorithms necessary. The scope and purpose of this thesis is to re-evaluate the embedding algorithms in the updated software environment and to study sources of systematic errors in the embedding procedure.

Chapter 2 summarises the theoretical aspects of the Higgs boson. The introduction to the particle accelerator Large Hadron Collider and the detector experiment Compact Muon Solenoid is followed by a summary of the Higgs boson discovery and a summary of the $H \to \tau\tau$ analysis, one of the first analyses that used the embedding procedure to suppress the $Z \to \tau\tau$ background within the Compact Muon Solenoid experiment. Chapter 3 explains the idea behind the embedding procedure and the different algorithms used for its implementation. In Chapter 4 the algorithms are validated using the muon embedding. In Chapter 5, the embedding of τ-leptons is studied in the example of the final state where one of the two τ-leptons decays leptonically into a muon and the other τ-lepton decays hadronically. Finally, Chapter 6 presents a summary of the results of this thesis and gives an outlook on possible future studies of the embedding.

2 Higgs Physics at the LHC

The Standard Model of particle physics (SM) describes fundamental particles and their interactions. It unifies the description of all discovered fundamental particles and the electromagnetic, weak and strong interaction in one single theory. A existence of massive gauge bosons and the preservation of the principles of local gauge symmetry are realised in the Standard Model with the Brout-Englert-Higgs mechanism.

This mechanism and the role of the Higgs boson therein are explained in Chapter 2.1 The Large Hadron Collider (LHC) together with the detectors *A Toroidal LHC ApparatuS* (ATLAS) and *Compact Muon Solenoid* (CMS) located at two separate intersection points of the accelerator are outlined in Chapter 2.2. In Chapter 2.3, the discovery of the Higgs boson is summarised, followed by the CMS analysis for the search of the Higgs boson in the decay channel into two τ-leptons in Chapter 2.4.

2.1 The Role of the Higgs boson in the Standard Model of Particle Physics

The discovery of the Higgs boson was a milestones in the past decades of particle physics. Before its discovery the major question of how the principles of local gauge invariance can be preserved with massive particles was still unanswered. Related to this, it was also unclear, why the gauge bosons of the weak force have such large masses. The observable result of the large gauge boson mass is that the weak force is short ranged and weak at low energies while it has a coupling strength similar to the electromagnetic force at high energies.

The Standard Model of Particle Physics

The SM describes the electromagnetic, weak and strong force and their interaction with elementary particles. In a quantum field theory, fundamental forces and particles are described as fields, interactions are mediated by gauge bosons that couple to the respective charge of each force. The three forces in discussion are represented by a

$$SU(3)_c \times SU(2)_L \times U(1)_Y$$

symmetry in an external hyperspace. The strong force obeys an $SU(3)_c$ colour symmetry, resulting in eight massless gauge bosons, called gluons. The gluons carry a colour charge and therefore couple to colour charged particles like quarks but also to other gluons. This gives rise to phenomena like confinement and asymptotic freedom of the strong force. The $SU(2)_L$ is a symmetry in the space of weak isospin. Its

index L indicates that non-trivial $SU(2)$ transformations only act on the left-handed components of all particles. For each of the three generators of the $SU(2)$ symmetry, there is a corresponding gauge field $W_\mu^a, a = 1, 2, 3$. By defining operators that act on the Lagrangian density like ascending and descending operators from quantum mechanics, two of the three gauge fields can be rewritten as

$$W_\mu^+ = \frac{1}{\sqrt{2}}(W_\mu^1 - iW_\mu^2)$$
$$W_\mu^- = \frac{1}{\sqrt{2}}(W_\mu^1 + iW_\mu^2)$$

These correspond to charged current interactions of the weak force, mediated by the exchange of a charged W^+ and W^- boson. In this representation, the remaining gauge field, W_μ^3, does not contain all observed neutral current interactions, since it only couples to left-handed particles. The correct representation of neutral current interactions can be achieved by extending the $SU(2)_L$ symmetry by the $U(1)_Y$ symmetry that also acts on the right handed component of the fields. This introduces a new gauge field B_μ.

With this extension, the fields corresponding to the physical Z boson and photon field are derived from a rotation of the fields W_μ^3 and B_μ by the weak mixing angle θ_W as

$$Z_\mu = \cos(\theta_W)W_\mu^3 - \sin(\theta_W)B_\mu$$
$$A_\mu = \sin(\theta_W)W_\mu^3 + \cos(\theta_W)B_\mu$$

The weak mixing angle is derived from the coupling constants g and g' of the $SU(2)_L$ and $U(1)_Y$ symmetries as

$$\cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}} \quad \text{and} \quad \sin\theta_W = \frac{g'}{\sqrt{g^2 + g'^2}}$$

The Lagrangian density of the $SU(2)_L \times U(1)_Y$ symmetry describes the full structure of electroweak interactions. The weakness of this theory is that the mass of the weak gauge bosons break the local gauge invariance. Therefore, an additional mechanism is needed to explain the symmetry breaking and the existence of massive gauge bosons.

The Higgs boson in the Standard Model of Particle Physics

A mechanism that can explain the weak boson masses and restore local gauge symmetry was first proposed in 1964 by Peter Higgs, François Englert, Robert Brout and others [3–8]. A solution was found in the concept of spontaneous symmetry

breaking where the Lagrangian density is invariant under symmetry transformations and the symmetry is broken by a non zero energy ground state of the system.

The mechanism proposes the existence of a complex $SU(2)$ doublet field Φ with the potential $V(\Phi^\dagger\Phi)$ and the Lagrangian density $\mathcal{L}_{\text{Higgs}}$ as

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}$$
$$V(\Phi^\dagger\Phi) = -\mu^2\Phi^\dagger\Phi + \lambda\left(\Phi^\dagger\Phi\right)^2$$
$$\mathcal{L}_{\text{Higgs}} = (\partial_\mu\Phi^\dagger)(\partial^\mu\Phi) - V(\Phi).$$

With the vacuum expectation value $v = \sqrt{\frac{\mu^2}{2\lambda}}$, the potential V has its minimum at the energy ground state

$$\Phi = \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.1}$$

This non zero ground state is the cause of the spontaneous breaking of the electroweak gauge symmetry.

A one dimensional illustration of this potential is shown in Figure 2.1.

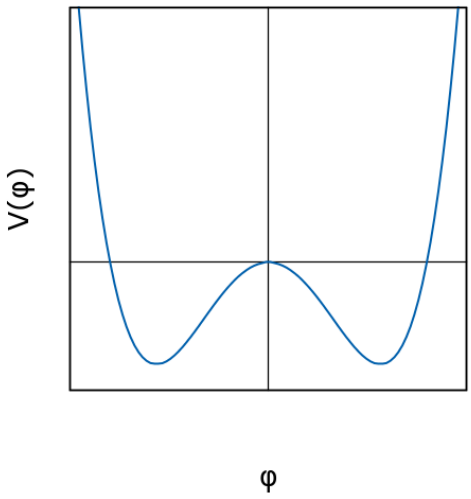


Figure 2.1: One dimensional illustration of the potential V of the Higgs doublet field Φ. The states in the minima correspond to the non vanishing energy ground state of the potential v. This causes the spontaneous symmetry breaking.

Three of the four degrees of freedom of the field Φ are absorbed by the masses of the weak gauge bosons. The remaining degree of freedom turns into the Higgs field H.

Weak gauge boson masses

When imposing local gauge invariance by performing the transition from the partial derivative ∂_μ to the covariant derivative

$$D_\mu = \partial_\mu - ig'\frac{Y_\Phi}{2}B_\mu - \frac{i}{2}gt^aW_\mu^a \qquad a = 1,2,3$$

the masses of the gauge bosons can be derived from the ground state of the Higgs field as given in equation 2.1. The Higgs field H is then introduced by the expansion in its energy ground state

$$\Phi = \begin{pmatrix} 0 \\ v + \frac{H}{\sqrt{2}} \end{pmatrix}.$$

The kinetic term of the Lagrangian density $\mathcal{L}_{\text{Higgs}}$ becomes

$$D_\mu\Phi^\dagger D^\mu\Phi = \frac{1}{2}\partial_\mu H\partial^\mu H + \frac{g^2+g'^2}{4}\left(v+\frac{H}{\sqrt{2}}\right)^2 Z_\mu Z^\mu + \frac{g^2}{4}\left(v+\frac{H}{\sqrt{2}}\right)^2 W_\mu^+W^{\mu-}$$

The mass terms of the gauge fields W_μ^+, W_μ^- and Z_μ are generated by the coupling to the vacuum expectation value v as

$$\left(\frac{g}{2}\right)^2 v^2 W_\mu^+W^{\mu-} \equiv m_W^2 W_\mu^+W^{\mu-}$$
$$\frac{g^2+g'^2}{4}v^2 Z_\mu Z^\mu \equiv m_Z^2 Z_\mu Z^\mu$$

From the Lagrangian density, the coupling of the Higgs boson to the weak gauge bosons can be read off to be

$$f_{H\to VV} = i\frac{2m_V^2}{v} \tag{2.2}$$

(2.3)

The numeric value of the vacuum expectation value v of the Higgs field can be determined from the precise measurement of the Fermi constant G_F from the lifetime of muons [9] in combination with the relation for the W boson mass as given above. From this follows

$$v = \frac{1}{\sqrt{\sqrt{2}G_F}} = 246.22\,\text{GeV}.$$

Fermion masses

The violation of gauge symmetry due to the asymmetry between left- and right-handed lepton masses can be overcome by a Yukawa coupling between the Higgs field and massless fermion fields. The corresponding gauge invariant Lagrangian density $\mathcal{L}_{\mathrm{Y,l}}$ for leptons is given by

$$\mathcal{L}_{\mathrm{Y,\,l}} = \mathcal{G}_{\mathrm{Y,\,l}}(\bar{l}_R \Phi^\dagger L_L + \overline{L}_L \Phi l_R) \qquad\qquad L_L = \begin{pmatrix} \nu \\ l_L \end{pmatrix}.$$

where $\mathcal{G}_{\mathrm{Y,\,l}}$ is the coupling strength, L_L the $SU(2)_L$ doublet of left handed leptons and l_R the $U(1)_Y$ wave function of right handed leptons. In the ground state of the field Φ the Lagrangian density becomes

$$\mathcal{L}_{\mathrm{Y,\,l}} = \frac{v}{\sqrt{2}} \mathcal{G}_{\mathrm{Y,\,l}}(\bar{l}_R l_L + \bar{l}_L l_R) = m_l \bar{l} l$$

The masses for down type quarks can be generated equivalently with the corresponding $SU(2)_L$ doublet of left handed quarks and the $U(1)_Y$ wave function of the right handed down type quark

$$\mathcal{L}_{\mathrm{Y,\,d}} = \frac{v}{\sqrt{2}} \mathcal{G}_{\mathrm{Y,\,d}}(\overline{d}_R d_L + \overline{d}_L d_R) = m_d \overline{d} d$$

The mass terms for up-type fermions are derived when adding equivalent Yukawa coupling terms to the Lagrangian density with the charge conjugate Φ_c of the $SU(2)_L$ field Φ to the Lagrangian density for leptons as given above. This does not apply to the leptons though, because the as massless assumed up-type neutrinos do not couple to Φ. Unlike to the gauge bosons, the Higgs field couples to fermions via a Yukawa coupling and the coupling strength is linear proportional to the fermion mass as

$$f_{H \to ff} = i \frac{m_f}{v}. \tag{2.4}$$

(2.5)

2.2 The Large Hadron Collider and the Compact Muon Solenoid

In this section, the main particle accelerators and detectors located at CERN are described, laying special emphasis on the particle accelerator LHC and the detector CMS.

2.2.1 The Large Hadron Collider

The *Large Hadron Collider* (LHC) is the world's largest and most powerful particle accelerator. It is located underneath the premises and surroundings of the *Conseil Européen pour la Recherche Nucléaire* (CERN) near Geneva in Switzerland. Built in the 27 km long ring tunnel of its predecessor the *Large Electron Positron Collider* (LEP), the LHC is only the last step in a series of accelerators, designed to bring particles to energies that have never been reached before.

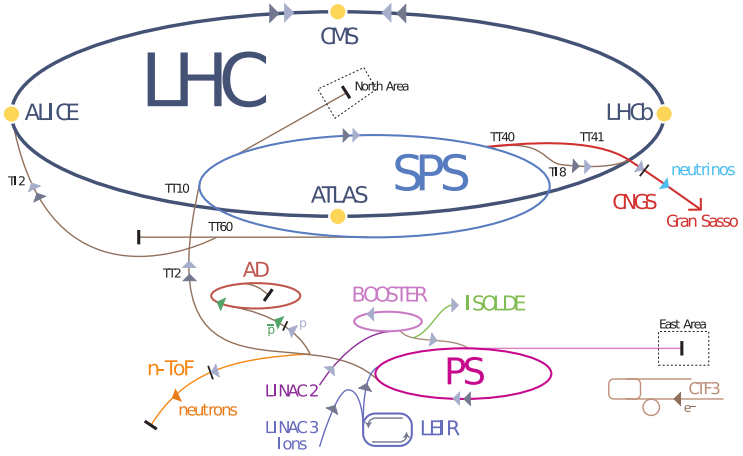


Figure 2.2: The complex of running accelerators at CERN. Protons are injected into the LHC from a series of smaller pre-accelerators [10].

The chain of acceleration as shown in Figure 2.2, starts at the linear accelerator Linac2 which brings the protons to energies of 50 MeV. The particles are then being injected into the PS Booster where they are accelerated to 1.4 GeV before they enter the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) which bring the protons to 25 GeV and 450 GeV respectively. From the SPS, the protons are finally injected into the LHC where they are accelerated up to 7 TeV. In the LHC, the proton beams are counter-circulating in two separate pipes which allows to collide them with a centre-of-mass energy of up to 14 TeV.

The particles in each pipe are being accelerated by eight radio frequency cavities along the tunnel. As particles get closer to the speed of light, these bunches stabilise as faster than average particles circulate on slightly larger radiuses and therefore need a longer time for each circulation. In the same way, slower than average particles need less time for each circulation as they run on slightly smaller radiuses.

To keep the charged particles within the curved beam pipe, they are deflected by a system of about 9600 magnets. The 1232 main dipole magnets are bending the particles on a quasi circular trajectory while the 392 main quadrupole magnets keep the beam focused. The dipole magnets are one of the most challenging instruments of the whole machine since their magnetic field strength is the limiting factor to the maximum reachable beam energy. Their high magnetic field strength of up to 8.3 T is

reached by using superconducting magnets. The last key ingredient to the accelerator is the ultra high vacuum of 10^{-13} atm to avoid collisions with air molecules within the pipes. The magnet system as well as the radio frequency cavities are operated at temperatures below 5 K that are reached by a liquid helium cooling system. With a total number of 2808 bunches of each 1.1×10^{11} protons, the LHC has a design luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ [11].

The first data taking period (run I) of the LHC took place from 30 March 2010 to 13 February 2013 and was performed at centre-of-mass energies of 7 TeV to 8 TeV. The second data taking period (run II) of the LHC has started on 3 June 2015 with an energy of 6.5 TeV per beam, resulting in a centre-of-mass energy of 13 TeV.

The bunches of oppositely running beams are brought to collision at the locations of the four main experiments ALICE, ATLAS, CMS and LHCb along the circle as illustrated in Figure 2.2. The two largest experiments ATLAS and CMS are general-purpose detectors, designed to be able to search for the Higgs boson as well as dark matter and super symmetry. The ALICE collaboration is studying the quark-gluon plasma that is believed to have existed a few microseconds after the Big Bang. For this, lead ions instead of protons are being collided at a centre-of-mass energy of 2.76 TeV per nucleon. The LHCb experiment is specialised in b-quark physics. The collaboration is studying CP violation in the sector of b-hadrons measured with a single arm forward spectrometer. One of their main goals is to reach a better understanding of the matter-antimatter asymmetry in our Universe.

2.2.2 The Compact Muon Solenoid

The *Compact Muon Solenoid* (CMS) is one of the two large, general-purpose detectors located along the LHC. The main requirements to the detector were the ability to identify and distinguish between all particles with lifetimes larger than a few nanoseconds. For this purpose, it was build around a 7 m$\times$ 13 m large superconducting solenoid magnet producing an homogeneous magnetic field strength of 3.8 T on its inside. Its size allowed for the tracker and the calorimeters to be built inside of the solenoid. The large magnet with a high field strength is important to measure the transverse momentum from the curvature of the track of charged particles. The energy of particles is measured with the calorimeters. An overview over the detector and its components is given in Figure 2.3.

Its main active components from the beam pipe to the outside are the tracker to identify charged particles, the electromagnetic calorimeter (ECAL) to detect electromagnetically interacting particles, the hadronic calorimeter (HCAL) to detect strongly interacting particles and on the outside of the solenoid the muon chambers that contribute to the identification of muons. All components consist of a barrel shaped structure along the beam pipe complemented by endcaps on each side to achieve an almost complete coverage of the space around the primary interaction point.

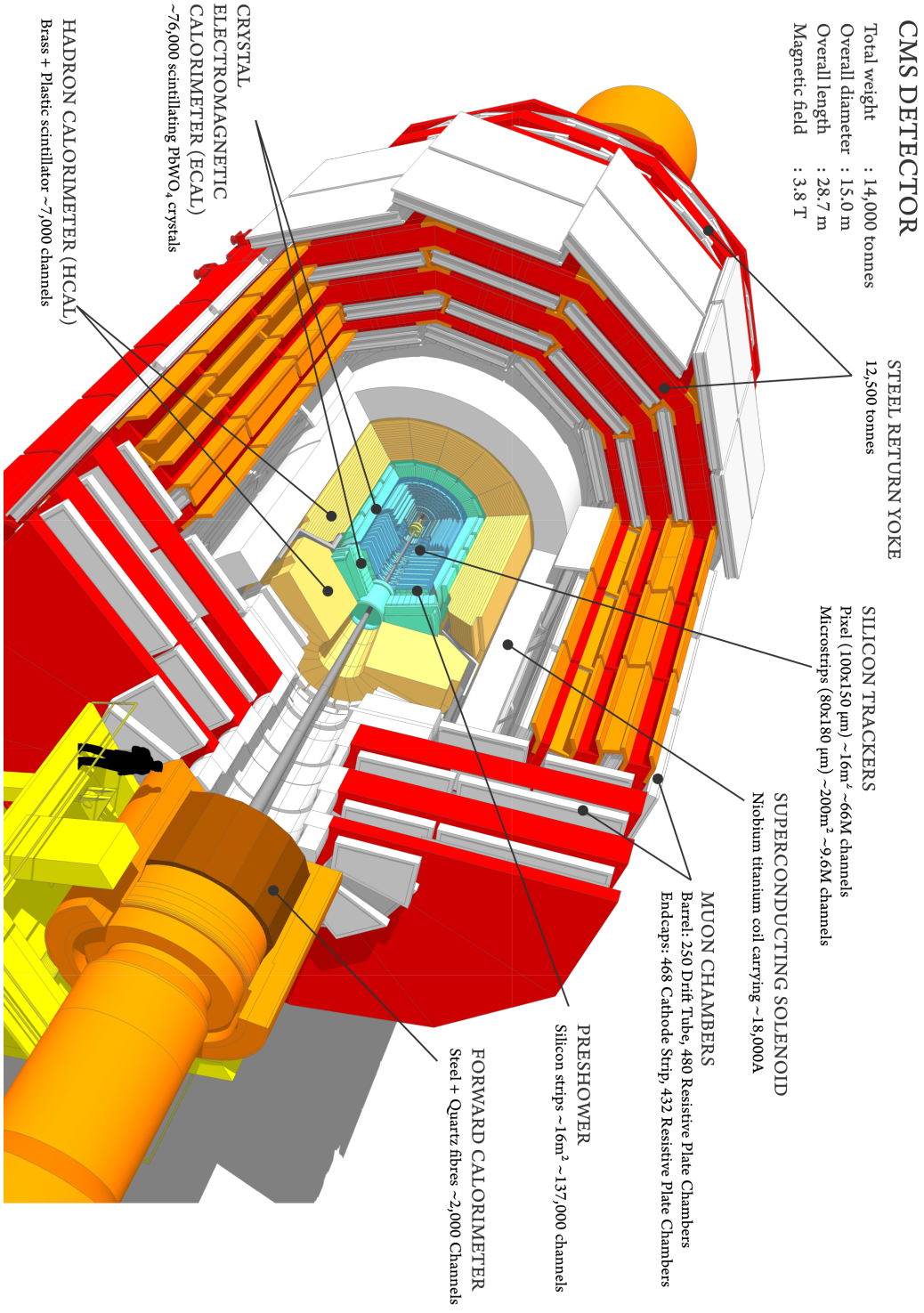
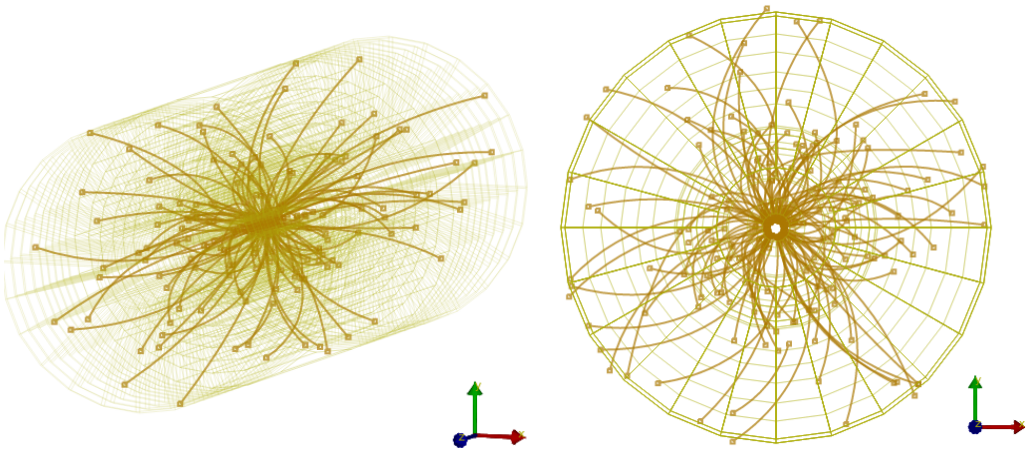


Figure 2.3: View of the main sections and components of the CMS detector [12].

The Inner Track Detector

The inner track detector records the trajectories of charged particles close to the collision vertex of the initially colliding protons. Due to its location directly around the beam pipe, it can also track the path of short lived and short ranged particles like B mesons. Its main purpose is to provide data to determine decay vertices of those fast decaying particles and to determine the momentum of charged particles. The latter is done by measuring the curvature of the track as the particles pass several layers of the pixel and strip silicon tracker. An example of tracks reconstructed for one event is shown in Figure 2.4. The impact parameter resolution for tracks with high momentum is around 10 µm. The resolution is small enough to reconstruct the decay vertices of hadronic τ-lepton decays that are typically several 10 µm away from the primary collision vertex. With this track resolution, a momentum resolution of 0.7 % for a particle with a momentum of 1 GeV and 5 % for a particle with a momentum of 1 TeV can be achieved [13].



(a) **(b)**

Figure 2.4: Tracker model with the reconstructed tracks from one event. Figure 2.4a shows a three dimensional model of the tracker with the reconstructed tracks from one event. The curvature of these tracks can be seen best in the side view of the tracker in Figure 2.4b. The reconstructed tracks and their curvature are the key ingredients in reconstructing the momentum of traversing particles.

The Electromagnetic Calorimeter

The second sub-detector that particles from the primary interaction pass through is the electromagnetic calorimeter (ECAL). This detector part was built to reconstruct the energy of electromagnetically interacting particles like electrons and photons. Since the ECAL and the hadronic calorimeter (HCAL) were built inside the solenoid,

dense materials had to be used to achieve short radiation lengths and therefore the name giving *compact* design. As detector material for the ECAL, crystals of lead tungstate were used. These have the advantage of having a short radiation length of only $X_0 = 0.89\,\text{cm}$. Therefore, large amounts of energy from electrons and photons are usually deposited in relatively small path lengths. Another advantage of lead tungstate is its quick response time. 80 % of the scintillation light is emitted within the 25 ns of two consecutive collisions. The disadvantages of this material are the low amount of scintillation light that is emitted when particles pass through and its high temperature sensitivity. Due to the chosen scintillator material, the temperature of the ECAL has to be kept stable within 0.1 K throughout the calorimeter.

The energy resolution of the ECAL barrel for electrons in test beams was determined to be [14]

$$\frac{\sigma_E}{E} = \frac{2.8\,\%}{\sqrt{E(\text{GeV})}} \oplus \frac{12\,\%}{E(\text{GeV})} \oplus 0.3\,\%. \tag{2.6}$$

From the calibration with a centre-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$, the energy resolution for the decay of a Z boson into electrons was determined to be 2 % in the central region of the barrel for pseudorapidities of $|\eta| < 0.8$ and 2 % to 5 % in the rest of the ECAL. For photons from a 125 GeV Higgs boson decay, the energy resolution was determined to be between 1.1 % and 2.6 % in the barrel and 2.2 % to 5 % in the endcaps. The absolute energy from $\text{Z} \to ee$ decays could be determined to a precision of 0.4 % in the barrel and 0.8 % in the endcaps [15].

The Hadronic Calorimeter

The HCAL is the last detector part on the inside of the solenoid. It was designed to measure the energy of hadronically interacting particles such as charged pions and kaons. The HCAL is a sampling calorimeter meaning it is made of alternating layers of absorber and scintillator material. As particles pass through the brass of the absorber, secondary charged particles and photons are created that produce scintillation light as they pass through the adjacent layer of plastic scintillator. The HCAL has a depth of 5.8 to 10 nuclear interaction lengths λ_{int}.

While the energy resolution of a monoenergetic beam of pions is between $\frac{22\,\%}{\sqrt{E}}$ and $\frac{10\,\%}{\sqrt{E}}$ for 30 GeV and 300 GeV pions respectively [16], the overall energy resolution of the HCAL is of the order of $\frac{100\,\%}{\sqrt{E}}$ [17]. The reason for the worsened energy resolution is the uneven response of the calorimeter to pions and electrons of the same energy. The origin of this unequal response are the different cross sections of electrons and pions. This is also expressed in the typically approximately 10 times smaller radiation length X_0 of electromagnetic interacting particles compared to the nuclear interaction length λ_{int} for strongly interacting particles. The ratio of energy deposited by an electron and a pion of the same energy is approximately $\frac{e}{h} = 1.4$. Since this difference is not compensated for and the ratio $\frac{e}{h} \not\approx 1$, the calorimeter is a so called non-compensating calorimeter.

The Muon Chambers

The outermost part of the detector contains the muon chambers. They are designed to identify muons by measuring energy deposits from gas ionisation. Muons e.g. from Z boson decays occur with energies where most of them are minimal ionising particles and have a relatively long mean free path. This allows for them to be detected several meters away from the primary collision point in an area of the detector where almost no other particles punch through. The muon chambers are a combination of gas ionisation detectors in the form of drift tubes for precise position measurement and resistive plate chambers for fast trigger information. The drift tubes are interleaved with the return yoke of the magnet coil that also acts as absorber for other particles than muons. Considering the approximately 1 to 2 nuclear interaction length of the ECAL and the 5.8 to 10 nuclear interaction lengths of the HCAL, up to 0.1 % of high energetic hadrons are expected to punch through to the first layer of the muon chambers. Therefore, only particles that pass through several muon chambers are identified as muons. By measuring their tracks in multiple layers of drift tubes combined with the information of the inner track detector, the curvature of the track and thereby the momentum of the muons is reconstructed.

At $\sqrt{s} = 7\,\text{TeV}$, muons of energies of more than approximately 8 GeV could be reconstructed with an efficiency of more than 95 % within the covered area of pseudorapidity $|\eta| < 2.4$. Muons from Z boson decays can be reconstructed with a precision of 0.2 % regarding their overall momentum. The transverse momentum, p_T, of muons below 100 GeV can be measured with a precision of 1 % to 6 % depending on the pseudorapidity. With data from cosmic muons, the transverse momentum resolution could be determined to be better than 10 % in the central region of the detector for muons of transverse momentum up to 1 TeV [18].

Particle Reconstruction

The particle reconstruction in CMS is done with the so called *particle flow* algorithm[19]. The idea is, to follow the traces that a particle leaves in the detector and use the combined information of all sub-detectors involved to identify the particle and to reconstruct its kinematics. This technique can be used in CMS due to the high granularity and very good resolution of the ECAL as well as the excellent inner track detector.

The algorithm starts with extrapolating the track of a particle as reconstructed from the inner track detector. Energy deposited in the calorimeter cells along this track is then combined with this track. To prevent the overlapping from close by showers, a clustering of the energy cells of the calorimeter is applied. Once an energy cluster has been attributed to a certain track, the cluster is removed from the remaining event to not attribute energy several times to different particles. Using this information in combination with the event topology and signal from the muon chambers, muons, electrons and charged hadrons are identified. The remainder of the events are then energy deposits that cannot be linked to a track. These are

contributed to photons for energy deposits in the ECAL and neutral hadrons for energy deposits in the HCAL.

The impact of the particle flow technique can be demonstrated easily by looking e.g. at the reconstruction of jets in the detector. Using the particle flow technique, between 95 % and 97 % of the energy of jets up to 600 GeV can be reconstructed, compared to 60 % to 80 % using only the information from the calorimeter read out. This leads to a significant improvement to the energy resolution of jets as shown in Figure 2.5.

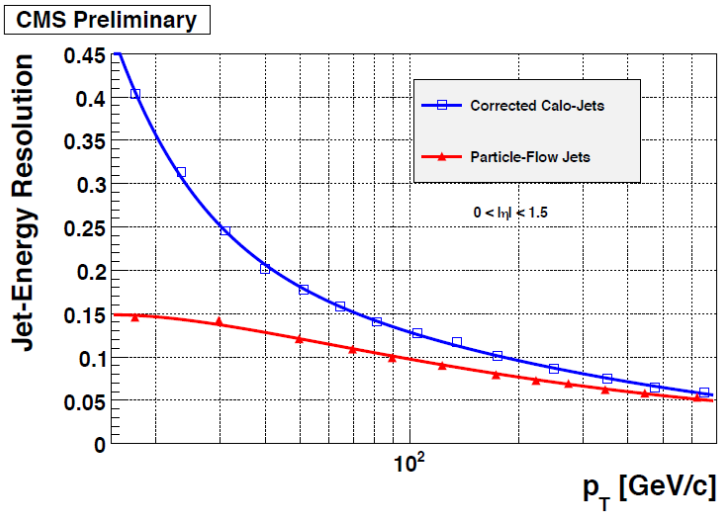


Figure 2.5: Energy resolution of jets reconstructed with the particle flow algorithm compared to jets reconstructed only with the calorimeter information. The particle flow reconstructed jets show an overall improved energy resolution, especially at low jet energies [19].

The algorithm provides an improved energy resolution, lower rates of misidentified particles and reduced systematic errors for all identified particles, significantly improving the overall performance of the CMS detector.

At the LHC, several inelastic proton-proton collisions take place at each bunch crossing. The average number were 9 in 2011, 21 in 2012 and are expected to be more than 40 in the beginning of run II of the LHC. Of those usually only one collision belongs to a hard inelastic scattering process. For the vertex of each of these proton-proton collisions, the sum of the squared momenta of tracks associated with it is calculated. Then, the vertex with the highest momentum is chosen to be the primary vertex. Tracks not associated with this vertex are referred to as pileup (PU). Due to finite response and read out times of the detector components, signal from previous and subsequent collisions also contributes to the signal of each event. These contributions are called out of time pileup and are suppressed by the reconstruction algorithms of CMS Software (CMSSW).

The reconstructed particles are combined to collections of electrons, muons and charged hadrons from the primary vertex, charged hadrons from pileup vertices, neutral hadrons and photons. The latter two collections contain particles from both the primary interaction vertex and the pileup interaction vertices. Since neutral hadrons and photons do not leave a track in the track detector they cannot be traced back to a specific vertex.

Using these particle collections, the isolation of leptons is calculated. The isolation estimates the amount of deposited energy from hadronic interactions close to a reconstructed lepton. If a large amount of energy from hadronic interactions is close to a reconstructed lepton, it is more likely that the reconstructed lepton is a misidentified charged hadron. Thus, by applying a selection on the isolation, the rate of misidentified leptons is reduced.

The isolation of leptons from the primary vertex is defined as

$$I^l = \sum_{charged\ hadr.} p_t + max\big(0, \sum_{neutral\ hadr.} p_t + \sum_{photons} p_t - \frac{1}{2} \sum_{charged\ hadr.,PU} p_t\big) \quad (2.7)$$

The sums therein are the scalar sum of transverse momentum of each particle collection within in a cone $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} \leqslant 0.4$, centred around the direction of the considered lepton. For hadronically decaying tau leptons, τ_h, particles that were used in the reconstruction of the τ_h candidate are not considered when calculating the isolation sum. The photon and neutral hadron collections also contain the signal from pileup for which they have to be corrected. Simulation has shown that approximately 1/3 of the energy of the hadronisation process in inelastic proton-proton scattering goes to photons and neutral hadrons whereas 2/3 go into charged hadrons. Therefore, the correction for the pileup contribution of the photons and neutral hadrons is taken from 1/2 of the scalar sum of transverse momentum from charged hadrons from pileup. The relative isolation is defined as $R^l = I^l / p_\mathrm{T}^l$.

Where the contribution of the individual components of the lepton isolation is studied, the scalar sums of transverse momentum from individual particle collections is calculated and taken as a measure for the isolation. Where only one of the components of the lepton isolation is used, the absolute and relative isolation are denoted with ch for the charged hadrons, nh for neutral hadrons, ph for photons and ch, PU for charged hadrons associated with pileup vertices. For example $I^{\mu,nh}$ represents the isolation of a muon only considering the sum of transverse momentum from neutral hadrons and $R^{\mu,nh}$ the corresponding relative isolation $I^{\mu,nh}/p_{\mathrm{T},\mu}$.

2.3 Discovery of the Higgs boson at the LHC

The main production mechanisms for the Higgs boson are via gluon fusion, vector boson fusion (VBF) and the associated production with a W or Z boson as shown in the leading-order Feynman diagrams in Figure 2.6. In the production via gluon fusion, only the Higgs boson is created. The Higgs boson can theoretically decay in any combination of particle and respective antiparticle, e.g. in a $\tau^+\tau^-$ pair. The

branching fractions of the different decay channels are only known once the mass of the Higgs boson is known. If produced via VBF, the Higgs boson is accompanied by two jets from the additional quarks in the process. In the production in association with a W or Z boson, up to two additional leptons and jets are generated.

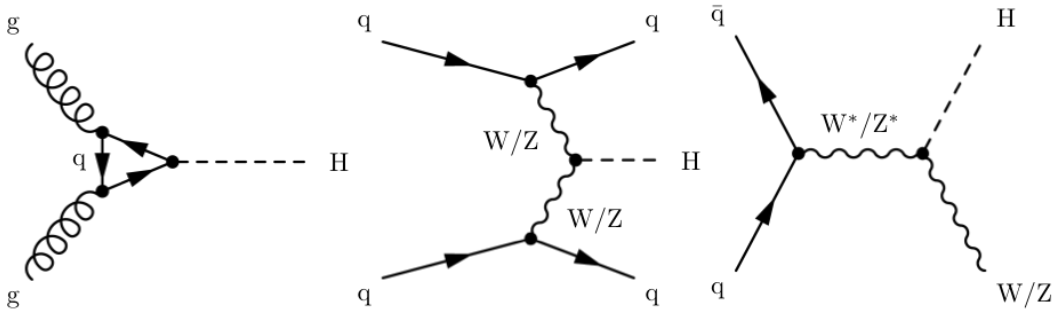


Figure 2.6: Leading-order Feynman diagrams of the main Higgs boson production mechanisms. From left to right the production via gluon fusion, vector boson fusion (VBF) and in association with a W or Z boson are shown [20].

On 4 July 2012, the discovery of a new boson with a mass around 126 GeV was announced at CERN [1, 2]. The new particle was discovered independently by the collaborations ATLAS and CMS by combining the signal from several decay channels. By this time, both collaborations had not excluded a Standard Model Higgs boson mass between 115 GeV and 130 GeV.

The ATLAS collaboration presented results based on $4.8\,\mathrm{fb}^{-1}$ of recorded data at $\sqrt{s} = 7\,\mathrm{TeV}$ combined with $5.9\,\mathrm{fb}^{-1}$ at $\sqrt{s} = 8\,\mathrm{TeV}$. In the $H \to \gamma\gamma$ channel, the search was conducted in an invariant mass range of 110 GeV to 150 GeV. Events with two isolated photons of energies larger than 40 GeV and 30 GeV respectively were selected. The main background in this decay channel is a continuously smoothly falling background, e.g. from π^0-decays, jets misidentified as photons and QCD diphoton production. An excess in the number of events at 126.5 GeV over this smoothly falling background expectation is visible in Figure 2.7. This excess corresponds to a local significance of 4.5 standard deviations, σ, or a p-value of 2×10^{-6}. This decay channel had the best mass resolution and therefore dominates the mass measurement in the combination of decay channels studied by the ATLAS collaboration.

The second most significant excess was observed in the $H \to ZZ$ decay channel. Here, events where both Z bosons decay into electrons or muons were selected. From this decay channel only a very small number of events is expected to pass the event selection. Nevertheless, it can contribute significantly, for it has a very low background rate, leading to a signal to background ratio (S/B) close to 1. Additionally, it provides a good mass resolution since the mass can be fully reconstructed in the event. The search here was performed for Higgs boson masses between 110 GeV and 600 GeV. In this decay channel, an excess in the number of events was found for masses around 125 GeV. The observed excess corresponds to a local significance of 3.4σ or a p-value

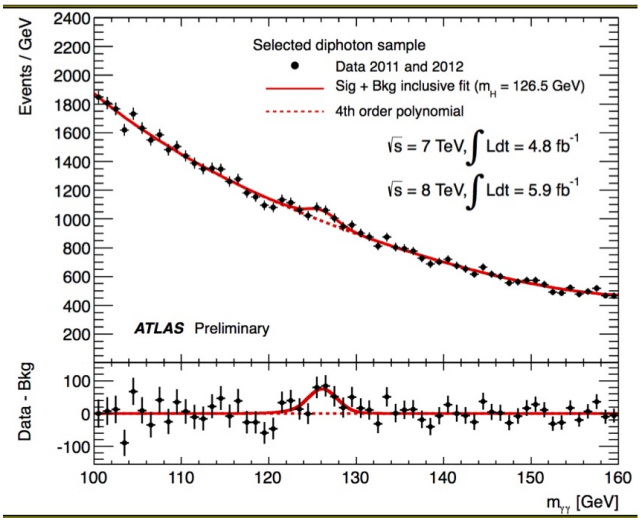


Figure 2.7: Reconstructed mass of selected di-photon events recorded with the ATLAS detector experiment. A significant excess over the background expectation is visible around masses of 126.5 GeV. This excess corresponds to a local significance of 4.5σ or a p-value of 2×10^{-6} [21].

of 3×10^{-4}. The combination of these two channels yields in a combined local significance of 5.0σ or a p-value of 3×10^{-7} for a mass of 126.5 GeV.

The CMS collaboration presented results from the decay channels $H \to WW$ and $H \to ZZ$ for masses between 110 GeV and 550 GeV and from the decay channels $H \to \gamma\gamma$, $H \to \tau\tau$ and $H \to b\bar{b}$ for masses below ≈ 150 GeV. For this analysis, $5.1\,\text{fb}^{-1}$ of recorded data at $\sqrt{s} = 7$ TeV and $5.3\,\text{fb}^{-1}$ at $\sqrt{s} = 8$ TeV were used.

In the $H \to \gamma\gamma$ channel a multivariate analysis was used, splitting events in 4 event classes based on a di-photon MVA output and two di-jet categories. This improved the sensitivity of this analysis by $\approx 15\,\%$ compared to a cut-based analysis. From the combination of all MVA categories in the analysis, a combined local significance of 4.1σ or a p-value of 2×10^{-5} was achieved for a mass of 125 GeV.

In the $H \to ZZ$ decay channel, a search for events where both Z bosons decayed into electrons or muons was performed. The main challenge here was the need for highest possible reconstruction efficiencies since a low number of events in this decay channel was expected. The irreducible backgrounds of $qq \to \text{ZZ} \to 4l$ and $gg \to \text{ZZ} \to 4l$ from quark- and gluon-fusion were estimated using Monte Carlo simulation. The reducible backgrounds were estimated from the extrapolation of control samples using data. With this a local significance of 3.2σ was achieved for a mass of 125.5 GeV.

The other in CMS examined decay channels $H \to WW$, $H \to \tau\tau$ and $H \to b\bar{b}$ did not yet have significant excesses above 3σ by this time. They were taken into consideration for a combined result from all channels. The profile plot of the p-value over the Higgs boson mass from this combination is shown in Figure 2.8. The achieved combined significance was 4.9σ or a p-value of 5×10^{-7} for a mass of 125.3 ± 0.6 GeV.

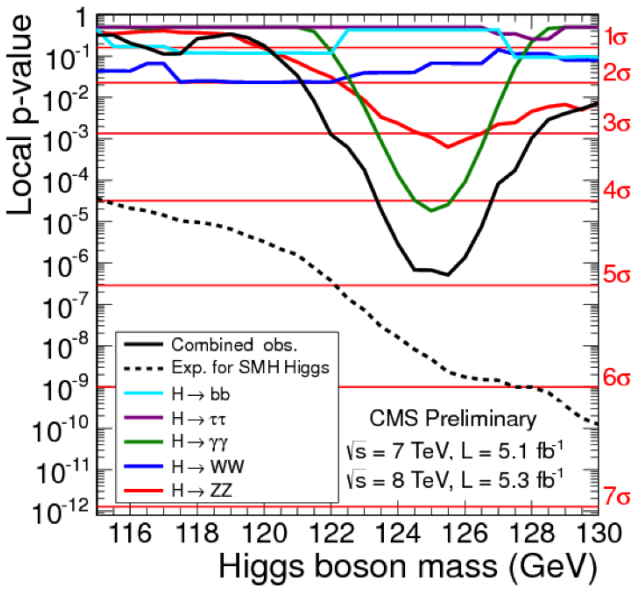


Figure 2.8: The local p-value over the Higgs boson mass as presented by CMS on 4 July 2012. The individual studied Higgs boson decay channels as well as their combination are shown. The combined significance of the extracted signals is 4.9σ, corresponding to a p-value of 5×10^{-7} at a mass of $125.3 \pm 0.6\,\text{GeV}$ [22].

With both experiments independently observing an excess in all examined decay channels and the combined significances of 5.0σ and 4.9σ, the discovery of a SM Higgs boson-like particle was announced. To prove that this boson was the long searched Higgs boson, more decay channels and properties of the discovered particle had to be studied. One missing link was the predicted coupling of the Higgs boson to leptons via the Yukawa coupling. The first evidence for this was given by the CMS collaboration in the $H \to \tau\tau$ decay channel, which had been analysed based on the complete dataset of the first data taking period [20].

Additionally, the coupling structure of the newly discovered particle had to be tested. For this, coupling strength parameters κ_j were introduced to compare the measured coupling strengths with the SM prediction. The parameters are normalised such that the SM prediction corresponds to a value of $\kappa_j = 1$ for each parameter. The result of a maximum likelihood fit for the κ_j of all studied decay channels, as analysed by the CMS collaboration, is shown in Figure 2.9a. The black points represent the measured value for the coupling strength, the red and blue bars the 68 % and 95 % confidence level (CL) intervals. All measured coupling strengths are within the 68 % CL uncertainty. This confirms that the new particle has a SM Higgs boson-like coupling structure.

As introduced in equation 2.2 and 2.4, the coupling strength of the new boson has to be linear proportional to the mass of fermions and quadratically proportional to the mass of vector bosons. Figure 2.9b shows the determined coupling strength of

the Higgs boson as a function of the particle mass as determined based on the full dataset recorded in CMS during the first data taking period of the LHC. For this figure, the coupling strength of fermions and vector bosons have been transformed like

$$|f^{obs}_{H\to ff}| = \kappa_f \cdot |f^{SM}_{H\to ff}| \qquad = \kappa_f \cdot \frac{m_f}{v} \qquad f = \mu, \tau, b, t \tag{2.8}$$

$$\sqrt{\frac{|f^{obs}_{H\to VV}|}{2v}} = \sqrt{\kappa_V} \cdot \sqrt{\frac{|f^{SM}_{H\to VV}|}{2v}} \qquad = \sqrt{\kappa_v} \cdot \frac{m_V}{v} \qquad V = \mathrm{W}, \mathrm{Z} \tag{2.9}$$

to account for the different coupling structure of fermions and bosons.

A maximum likelihood fit of these transformed couplings was done on the modified parameters $\kappa_f = v \times \frac{m_f^{\epsilon}}{M^{\epsilon+1}}$ and $\kappa_V = v \times \frac{m_f^{2\epsilon}}{M^{2\epsilon+1}}$. The parameter M therein is the fit parameter for the the vacuum expectation value v, the parameter ϵ is introduced to account for possible deviations from the expected linear behaviour. The most probable fit parameters were $M = 245 \pm 15\mathrm{GeV}$ and $\epsilon = 0.01 \pm^{0.041}_{0.036}$. Both values are within the expectations for the vacuum expectation value and the linear behaviour of the transformed coupling strength. Therefore, the coupling of the new discovered particle is found to be as expected for a SM Higgs boson.

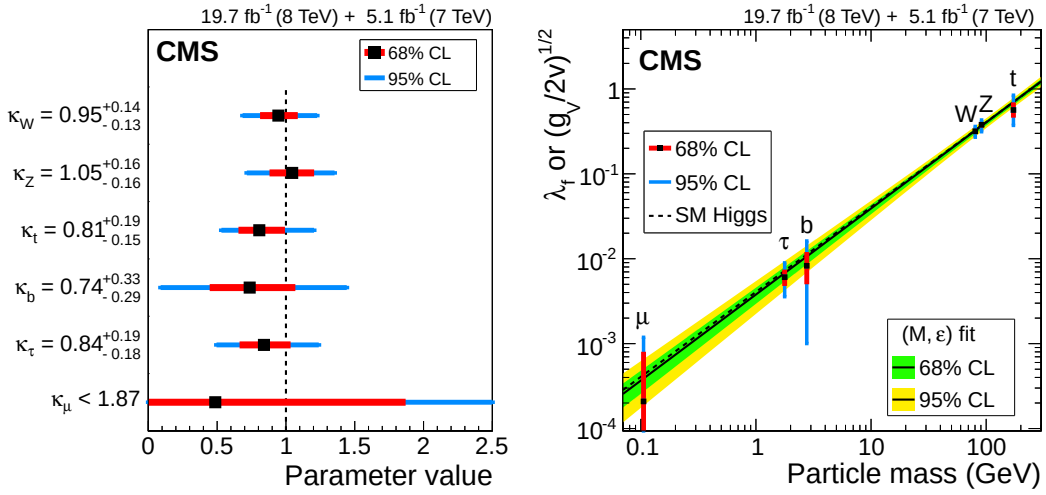


Figure 2.9: Figure 2.9a shows the maximum likelihood fit for the normalised coupling strength parameters κ_j. All parameters are within their 68 % confidence levels [23]. The interaction strength of the Higgs boson and different Standard Model particles as a function of the transformed parameters as given in equation 2.8 and 2.9 is shown in Figure 2.9b [23]. The fit is compatible with the linear dependency of the Higgs boson coupling strength to the particle mass expected with the used transformation.

2.4 The $H \to \tau\tau$ analysis

In June 2014, the CMS collaboration presented evidence for a Standard Model Higgs boson decaying into a pair of τ-leptons [20]. The analysis was performed on the dataset recorded during the first data taking period of the LHC. The dataset corresponds to an integrated luminosity of $4.9\,\text{fb}^{-1}$ at a centre-of-mass energy of 7 TeV and $19.7\,\text{fb}^{-1}$ at 8 TeV.

Where the Higgs boson is produced via gluon fusion, the final states with $H \to \tau\tau$ decays contain only two charged leptons. In case of the production via VBF, two additional jets are produced. Events with at least two reconstructed τ-leptons were split into six mutually exclusive datasets based on the reconstructed final states of the τ-leptons. Depending on whether each of the two τ-leptons decayed into an electron (e), muon (μ) or hadronically (τ_h), these final states were accordingly $ee, e\mu, e\tau_\text{h}, \mu\mu, \mu\tau_\text{h}$ and $\tau_\text{h}\tau_\text{h}$. All these channels have the Drell-Yan production of $\text{Z} \to \tau\tau$ decays as main irreducible background. The reduction of this background and its systematic uncertainties is the main goal of the so called *embedding procedure* as described in Chapter 3 and studied from Chapter 4 onwards.

The production in association with a W or Z boson results in final states with one or two additional leptons. These decay channels do not have the $\text{Z} \to \tau\tau$ background as large irreducible background. Also due to low event yields, the resulting decay channels did not contribute significantly to the result of the $H \to \tau\tau$ analysis. Therefore, these decay channels are not considered here any further.

The events were categorised so the number of selected electrons, muons and τ-leptons in the event corresponded to one of the 6 different decay channels, e.g. two muons in the $\mu\mu$ channel. A high level trigger (HLT) had to accept the event. The HLT requirements for each channel are given in Table 2.1. A requirement like '$\mu(18)$' for the HLT means, a muon of at least 18 GeV transverse momentum, p_T, must have been reconstructed by the HLT algorithm. Some of the trigger requirements regarding the transverse momentum of the triggered leptons had to be changed in 2012 to cope with the higher instantaneous luminosity of the collider when the centre-of-mass energy was raised to 8 TeV. Where this was necessary, the original and raised values of the p_T threshold of the trigger are given in the table, separated by a semicolon.

Additionally to the HLT requirement, each lepton had to pass kinematic requirements on the transverse momentum, p_T, as well as on the pseudorapidity, η. The selection values for each channel are also given in Table 2.1. To be selected as leptons and the primary collision vertex, other requirements were imposed on the distance of closest approach of the trajectories. It was required to not be larger than $d_z = 0.2\,\text{cm}$ in the direction of the beam pipe and not larger than $d_{xy} = 0.045\,\text{cm}$ in the transverse plane. The last element of the baseline event selection was the isolation requirement. For electrons and muons, a requirement on the relative isolation, R^l, for the τ_h candidates, a requirement on the absolute isolation, I^l, had to be fulfilled for the events to be selected.

Channel	HLT requirement	Lepton selection criteria		
$\mu\tau_{\mathrm{h}}$	$\mu(12;18)$ & $\tau_{\mathrm{h}}(10;20)$	$p_{\mathrm{t}}^{\mu} > 17;20$ $p_{\mathrm{t}}^{\tau_{\mathrm{h}}} > 30$	$\lvert\eta^{\mu}\rvert < 2.1$ $\lvert\eta^{\tau_{\mathrm{h}}}\rvert < 2.4$	$R^{\mu} < 0.1$ $I^{\tau_{\mathrm{h}}} < 1.5$
$e\tau_{\mathrm{h}}$	$e(15;22)$ & $\tau_{\mathrm{h}}(15;20)$	$p_{\mathrm{t}}^{e} > 20;24$ $p_{\mathrm{t}}^{\tau_{\mathrm{h}}} > 30$	$\lvert\eta^{e}\rvert < 2.1$ $\lvert\eta^{\tau_{\mathrm{h}}}\rvert < 2.4$	$R^{e} < 0.1$ $I^{\tau_{\mathrm{h}}} < 1.5$
$\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ (2012 only)	$\tau_{\mathrm{h}}(35)$ & $\tau_{\mathrm{h}}(35)$ $\tau_{\mathrm{h}}(30)$ & $\tau_{\mathrm{h}}(30)$ & jet(30)	$p_{\mathrm{t}}^{\tau_{\mathrm{h}}} > 45$	$\lvert\eta^{\tau_{\mathrm{h}}}\rvert < 2.1$	$I^{\tau_{\mathrm{h}}} < 1$
$e\mu$	$e(17)$ & $\mu(8)$ $e(8)$ & $\mu(17)$	$p_{\mathrm{t}}^{l_1} > 20$ $p_{\mathrm{t}}^{l_2} > 10$	$\lvert\eta^{\mu}\rvert < 2.1$ $\lvert\eta^{e}\rvert < 2.3$	$R^{l} < 0.1;0.15$
$\mu\mu$	$\mu(17)$ & $\mu(8)$	$p_{\mathrm{t}}^{\mu_1} > 20$ $p_{\mathrm{t}}^{\mu_2} > 10$	$\lvert\eta^{\mu_1}\rvert < 2.1$ $\lvert\eta^{\mu_2}\rvert < 2.4$	$R^{\mu} < 0.1$
ee	$e(17)$ & $e(8)$	$p_{\mathrm{t}}^{e_1} > 20$ $p_{\mathrm{t}}^{e_2} > 10$	$\lvert\eta^{e}\rvert < 2.3$	$R^{e} < 0.1;0.15$

Table 2.1: Selection criteria for the six main final states in the $H \to \tau\tau$ analysis. The indices 1 or 2 correspond to the leptons with the highest and second highest p_{T}. The values for p_{T} and I^{l} are given in GeV. Where two by a semicolon separated values are given, the first value corresponds to the selection for the $\sqrt{s} = 7\,\mathrm{TeV}$ dataset while the latter value corresponds to the selection for the $\sqrt{s} = 8\,\mathrm{TeV}$ dataset [20].

The signal in the $e\mu, e\tau_{\mathrm{h}}, \mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ decay channel was extracted from the reconstruction of the invariant mass of the τ-lepton pair, $m_{\tau\tau}$, which was used as an estimator of the mass of the parent boson. In the reconstruction, only a part m_{vis} of this energy is visible in the detector since some of the energy of the parent boson goes into neutrinos which escape from the experiment undetected. If the Higgs boson is created via VBF or gluon fusion, the τ-lepton decay is the only source of neutrinos. Therefore, the amount of missing transverse energy, E_{T}^{miss}, can be assumed to mainly come from the emitted neutrinos. A dedicated maximum likelihood based algorithm uses this assumption to calculate a more precise estimator for the invariant mass, $m_{\tau\tau}$, than the bare visible energy and missing transverse energy.

In the ee and $\mu\mu$ channel the signal was extracted from a multivariate discriminating variable D. This was built from the output of two Boosted Decision Trees (BDTs) based on kinematic variables of the di-lepton system, on the distance of closest approach between the leptons, the missing transverse energy vector, $\vec{E}_t^{miss}$, and in case of two additional jets in the event the di-jet mass, m_{jj}, and their distance in pseudorapidity, $|\Delta\eta_{jj}|$.

The selected events in each channel were split up in several mutually exclusive event categories, designed to increase the sensitivity for the search for the SM Higgs boson. These categories were e.g. based on the number of jets apart from the jets identified as hadronically decaying τ-leptons. In events with two additional

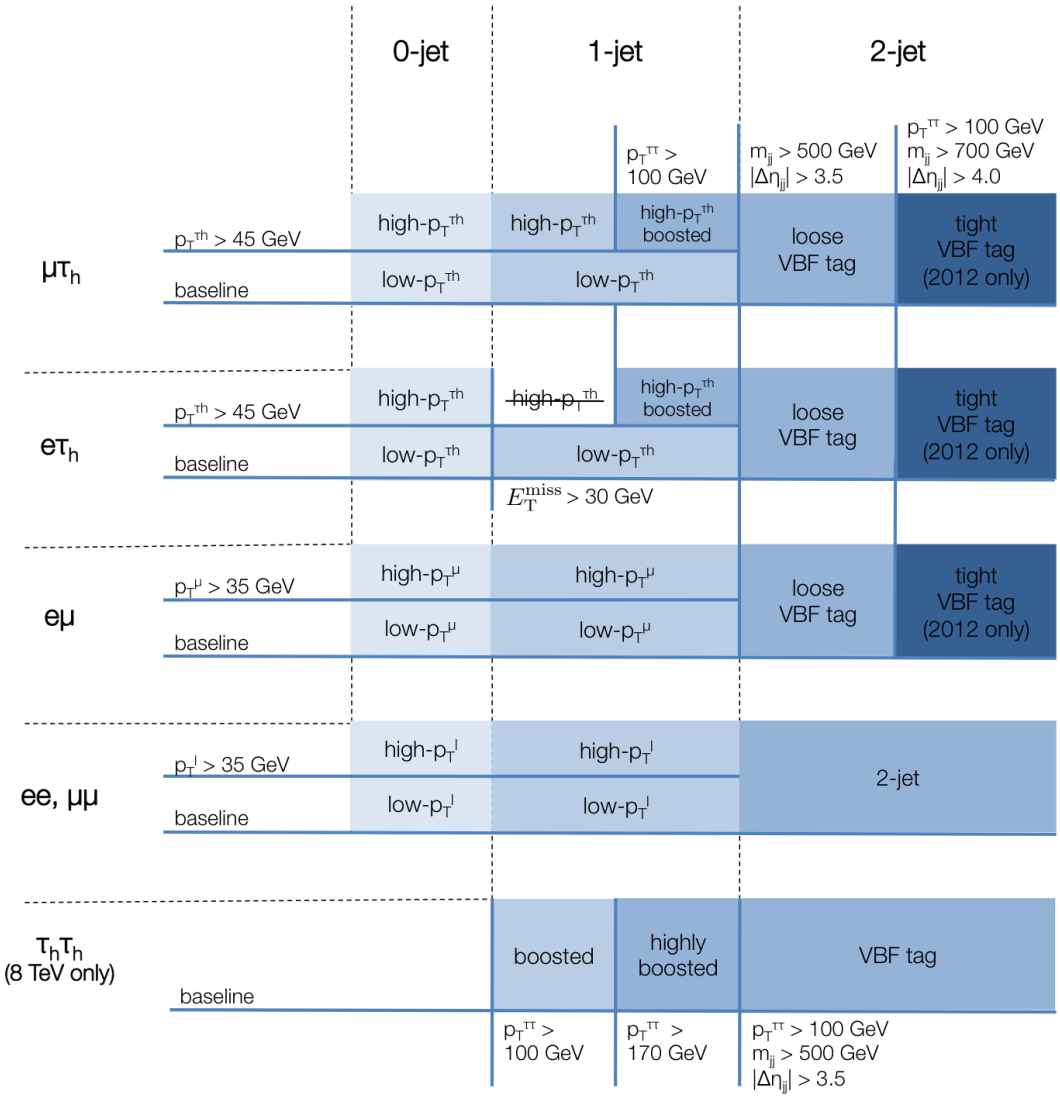


Figure 2.10: Sub categories of the main analysis in the six decay channels considered in the $H \to \tau\tau$ analysis. Each τ-pair decay channel is divided into a 0, 1 and 2 jet event category. Within these there are further sub categories based on the transverse momentum of the reconstructed leptons. In the 2 jet event category, a VBF tag is calculated for each event, increasing the discriminating power of the sub category [20].

reconstructed jets, an additional *VBF* tag was computed which was designed to select events where a Higgs boson has been produced via VBF. This tag is important for background suppression. It can e.g. suppress the $Z \to \tau\tau$ background, since in this decay, additional jets occure only rarely. Thereby, the discriminating power of the analysis is increased. Events without any additional jets were mainly used to constrain the $Z \to \tau\tau$ background since these events have a smaller discriminating power. The full set of event categories is shown in Figure 2.10.

The background estimation of the analysis is one example, where an embedding algorithm as described in Chapter 3 has been used to estimate the background from $Z \to \tau\tau$ events. Another important background to the $e\tau_h$ and $\mu\tau_h$ channels comes from events, where a W boson decayed leptonically and a jet was misidentified as a τ_h. The contribution from this background was taken from Monte Carlo simulation and normalised to the observed yield in a high m_T region where no other significant background than from W + jets events is expected. This contribution was then extrapolated to regions of lower m_T using simulation. The systematic uncertainty of this method was estimated to be between 10 % and 30 %, depending on the event category and decay channel.

The background from $t\bar{t}$ production is largest in the $e\mu$ channel and was estimated from simulation and normalised using a $t\bar{t}$-enriched control sample. The shape of this background was predicted by simulation and the yield adjusted to the observed yield in the control sample. The systematic uncertainty here was determined to be between 1.5 % and 7.4 %. The last considered background component is from QCD multijet events in the $e\tau_h$ and $\mu\tau_h$ channels. In these events, one jet can be misidentified as a τ_h and another jet as a lepton. This background was estimated from events where the two reconstructed leptons had the same charge. The amount of background from QCD in this sample was derived by subtraction the estimated yield from the previously mentioned background sources from the selected same charge events. The remaining events were assumed to be from the QCD background. The corresponding yield in the opposite charge signal sample is expected to be 1.06 times as large as the yield from the same charge events and was considered correspondingly. The systematic uncertainty of this background was assigned to be between 10 % and 50 %, depending on the considered $\tau\tau$ decay channel and event category.

Systematic uncertainties can be grouped in theory related uncertainties and uncertainties from experimental sources. Theory related uncertainties are mainly relevant for the estimation of the expected signal yields. Experimental sources of uncertainties arise from the reconstruction of physics objects and uncertainties in the background estimation. The most interesting systematic uncertainty in the scope of this thesis is for the background from $Z \to \tau\tau$ decays. The uncertainty of its yield is 3 % with an additional uncertainty on the extrapolation of the $Z \to \tau\tau$ mass distribution between 2 % and 14 % depending on the considered channel.

In Figure 2.11 the observed and predicted $m_{\tau\tau}$ distributions in the $\mu\tau_h$ and $e\tau_h$ channel for the 8 TeV dataset are shown. The tight VBF tagged subsets have low event yields, but the background contributions in this category, especially the $Z \to \tau\tau$

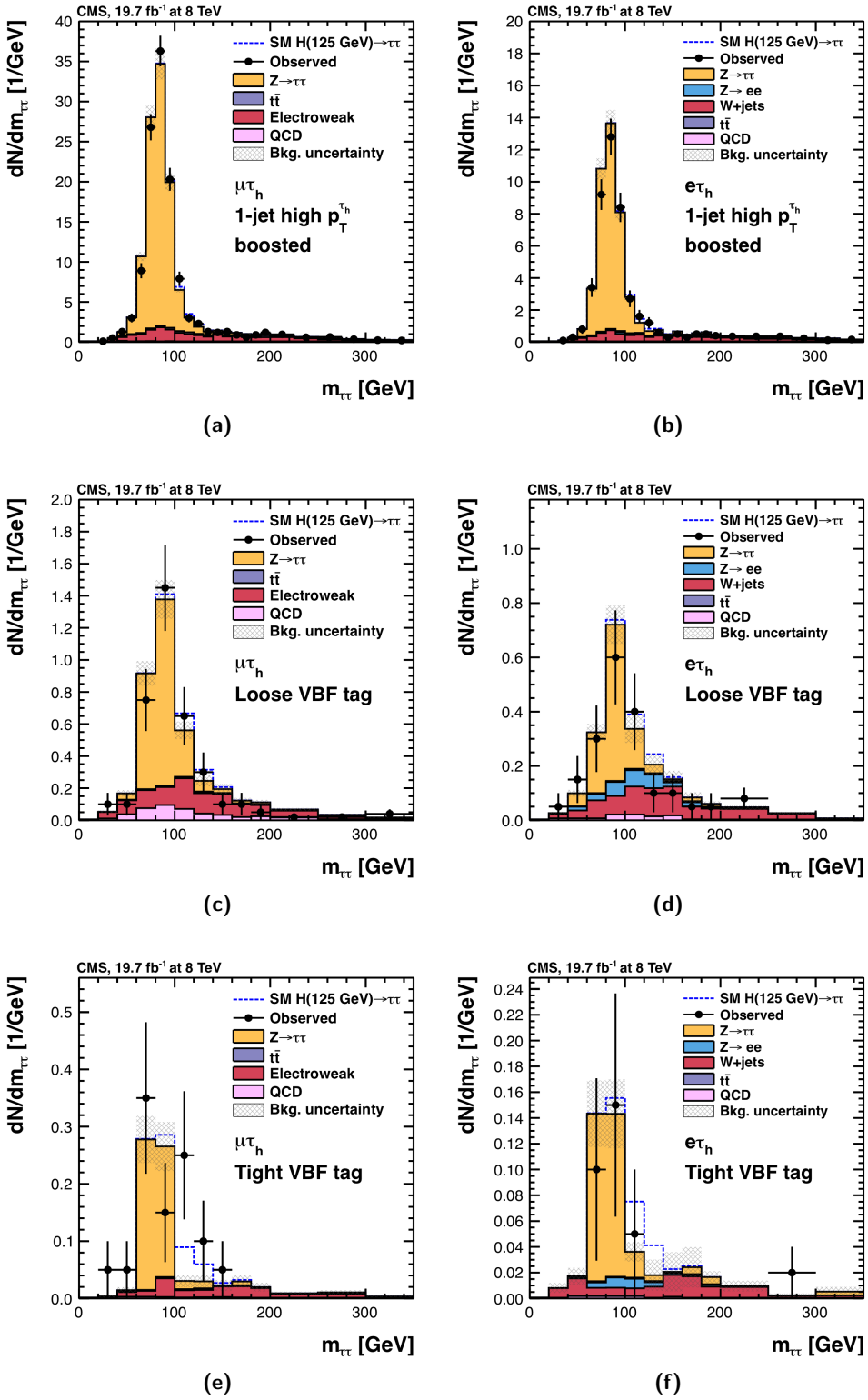


Figure 2.11: Observed and predicted $m_{\tau\tau}$ distributions in the $\mu\tau_{\mathrm{h}}$ and $e\tau_{\mathrm{h}}$ channel for the 8 TeV dataset. The plots show the distribution for the 1-jet high-$p_{\mathrm{T}}^{\tau_{\mathrm{h}}}$ boosted category (a, b), the loose VBF tag (c, d), and tight VBF tag (e, f). The expected background yields are shown as stacked histograms.

background, are strongly suppressed. Therefore, this category has a high S/B ratio and can contribute significantly with only a low number of selected events.

For the combination of the six described decay channels, the excess was quantified by calculating the local p-value as shown in Figure 2.12. This shows a significance larger than 3 standard deviations for the Higgs boson mass m_{H} between 110 GeV and 130 GeV with the maximum of 3.6 standard deviations at 125 GeV. The best fit value of the Standard Model signal cross section modifier μ was calculated to be 0.86 ± 0.29. This concludes evidence for a Standard Model Higgs boson decaying into two τ-leptons.

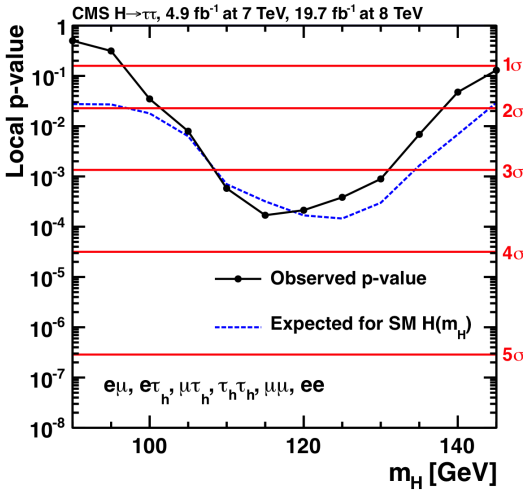


Figure 2.12: P-value of the observed excess over the Standard Model Higgs boson mass m_{H} in the six discussed decay channels. The observed excess corresponds to 3.6σ at 125 GeV [20].

3 The Embedding Procedure

With the large amount of data recorded at the experiments at the LHC, physicists are trying to discover rare processes by looking for small signals on top of large background contributions.

To distinguish between signal and backgroud processes, analysis strategies need to be developed to exploit all observable differences in event structures. Therefore, variables with a large discriminating power between signal and background need to be identified. The easiest way to extract a signal is by applying a selection on these discriminating variables.

Where bare requirements on single quantities are not sufficient anymore, multivariate analysis (MVA) strategies are used. These combine the information from all relevant variables in Boosted Decision Trees or Artificial Neural Networks to find the hyper plane with the largest discriminating power between signal and background. Using these methods, many background contributions can be reduced significantly, thereby increasing the sensitivity of the analyses. Background contributions which can be reduced by a set of discriminating variables are called *reducible*.

In some cases though, the event topologies are too similar, so no set of variables with a sufficient discriminating power can be found. One example is the $Z \to \tau\tau$ background in the $H \to \tau\tau$ analysis, introduced in Chapter 2.4. In proton-proton collisions, this background is caused by the Drell-Yan process $q\overline{q} \to Z/\gamma^* \to \tau\tau$, the annihilation of a quark-antiquark pair into a virtual Z boson or photon that then decays into two τ-leptons. Background contributions which cannot be suppressed effectively are called *irreducible* backgrounds. The only way to decrease the impact of an irreducible background on an analysis is to reduce its systematic uncertainties. By reducing the systematic uncertainties, better knowledge about the expected number of background events is achieved and therefore a significant excess can be identified earlier. Thus, methods to reduce the uncertainties of the irreducible $Z \to \tau\tau$ background need to be studied. One of these methods is the so called *embedding* procedure which is described in the following chapters.

The idea behind this procedure is explained in Chapter 3.1. The algorithms used for the embedding are illustrated in Chapter 3.2, followed by the selection process of events suitable for embedding that is described in Chapter 3.3. A method to reduce the main bias from this preselection is introduced in Chapter 3.4

3.1 The Embedding Idea

When estimating backgrounds, one distinguishes Monte Carlo (MC) simulation driven methods and data driven methods. Simulation driven methods estimate the contribution of a background from the theoretically expected number of events. Data driven methods are often used to estimate a background from signal free side bands of distributions.

The advantage of simulation driven methods is the theoretically arbitrary amount of events that can be simulated, reducing the relevant uncertainties to systematic ones. On the other hand, the detector response has to be simulated as well, adding additional uncertainties to the simulation driven background estimation like uncertainties on the jet energy resolution and noise in the read out of the detector. Another disadvantage of the simulation driven methods compared to data driven methods is that they can only depict the general level of understanding of physics. The main advantage of the data driven method is the perfect description of physics and the detector with the disadvantage of a limited number of events. Thus, data driven methods will benefit from reduced systematic errors at the cost of increased statistical errors.

Independently from the chosen data source, the estimation of the background from recorded $Z \to \tau\tau$ events has several sources for possible biases from the reconstruction of the two τ-leptons. The subsequent decays are difficult to reconstruct since the τ-leptons can either decay leptonically into an electron or a muon or hadronically into one or several hadrons. Due to the emission of the neutrinos and the energy resolution of the HCAL in CMS, important quantities like the invariant di-τ mass, $m_{\tau\tau}$, of the di-τ decay can only be reconstructed imperfectly.

This manifoldness of the τ-lepton decay eventually makes the selection of $Z \to \tau\tau$ events less efficient and less pure, thus increasing statistic and systematic uncertainties on the background estimation. This is where the *embedding* procedure comes into play. The procedure was introduced in the software framework of the CMS collaboration, CMS Software (CMSSW), in the search for a method to improve the background estimation from $Z \to \tau\tau$ decays.

The embedding is based on recorded $Z \to \mu\mu$ events which can be selected with very high efficiency and purity. Due to lepton universality, the decay of Z bosons into a pair of muons and τ-leptons is, apart from small deviations due to the different masses of muon and τ-lepton, identical. Therefore, besides the different Z boson decay products, these events show an identical event structure. Due to the smaller mass of the muons, the $H \to \mu\mu$ decay is suppressed compared to the $H \to \tau\tau$ decay by a factor of $\approx 3.5 \times 10^{-3}$. Thus, the selected $Z \to \mu\mu$ events can be considered free of a Higgs boson signal in the estimation of the $Z \to \tau\tau$ background in a $H \to \tau\tau$ analysis.

The embedding procedure works in several steps as sketched in Figure 3.1 and is described below:

1. **Select $Z \to \mu\mu$ events**
   Events with two well reconstructed and isolated muons are selected from the

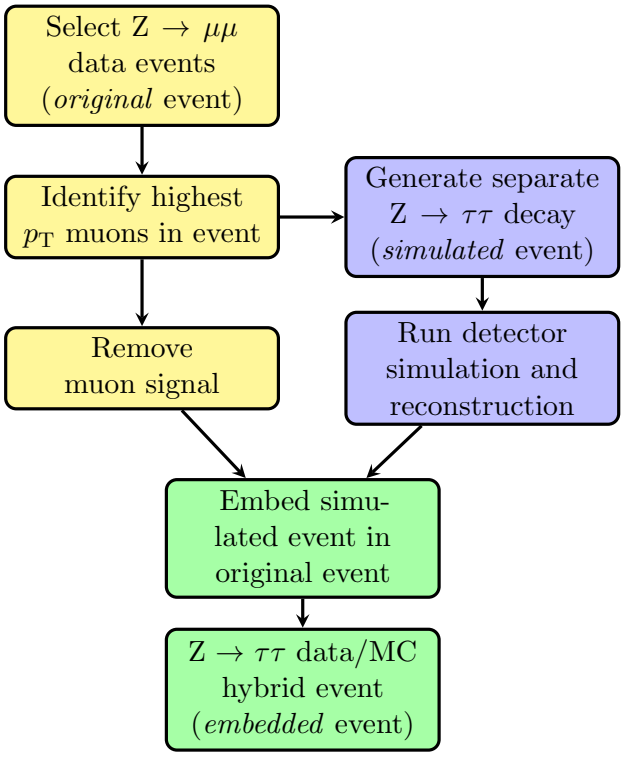


Figure 3.1: Scheme of the embedding procedure. The *original* $Z \to \mu\mu$ event is shown in yellow, the newly generated *simulated* event in blue and the merged *embedded* event in green. Starting from selected $Z \to \mu\mu$ events, the transverse momentum of the highest p_T muons in the original event is identified. The muons as well as calorimeter hits and tracks associated with them are then removed from the original event. Alongside, a new $Z \to \tau\tau$ MC event is generated, giving the τ-leptons in the event the exact position and the mass-corrected four-momentum of the removed muons from the original event. Then, the detector simulation and parts of the reconstruction algorithms are re-run, so the events can be merged, creating a $Z \to \tau\tau$ data/MC hybrid event.

recorded data. This di-muon selection and the used parameters are described in detail in Section 3.3. These selected $Z \to \mu\mu$ events build the foundation of the embedding and will be referred to as *original* event.

2. **Remove the muons from the original event**
   The calorimeter hits and tracks associated with the two reconstructed muons as well as the muons themselves are removed from the event.

3. **Simulate separate di-τ event**
   A separate event is generated that contains only a pair of simulated τ-leptons of the same kinematic properties as the two highest p_T muons from the original event. This event will be referred to as *simulated* event. The different mass of the τ-leptons is factored in by applying a small correction on the momentum

of the simulated τ-leptons, so the invariant mass, $m_{\tau\tau}$, of the created $Z \to \tau\tau$ event is identical to the reconstructed invariant mass, $m_{\mu\mu}$, from the original $Z \to \mu\mu$ event. Since the muons from the original event already emitted final state radiation and lost some of their energy, the emission of final state radiation is disabled for the embedded particles to not further smear out their kinematics. Afterwards, the detector simulation and parts of the reconstruction algorithms process this separate event.

4. **Embedding into the original event**
   The reconstructed τ-lepton decay products from the simulated event are then embedded into the remainder of the original $Z \to \mu\mu$ event, creating a data/MC hybrid event called *embedded* event.

Since the largest part of the event still originates from data, systematic uncertainties will be largely reduced in comparison to a fully MC simulated event.

Compared to a fully data driven background estimation, the embedded samples have a lower rate of misidentified $Z \to \tau\tau$ decays, since the event selection is based on the highly efficient muon identification. As a result, using embedding for the $Z \to \tau\tau$ decay, a larger number of well reconstructed events can be skimmed from the recorded data, reducing statistical uncertainties compared to a fully data driven background estimation in addition to reducing systematic uncertainties compared to a purely MC based method.

3.2 Embedding Algorithms

A crucial point when designing an embedding algorithm is to find a level of event reconstruction at which the merging of original and simulated event can be performed. The levels, where this can theoretically be done in a sensible way in CMSSW, are:

1. **Digitised detector output**
   The lowest possible level at which the merging can theoretically be performed is limited by the lowest level of simulation. This is the level of digitised detector output. The bandwidth when recording data on this level would be in the order of some hundred GB per second. This is some orders of magnitude larger than what the data storage infrastructure in CMS was designed and funded for, making it impossible to store this level of event reconstruction. Thus, the embedding at the level of digitised detector output is not possible due to computational limitations.

2. **Reconstructed Hits and Tracks (*Rec-Hit* Embedding)**
   The lowest level of event reconstruction where recorded data is available is the level of reconstructed calorimeter hits and tracks. The simulated event is also reconstructed to the level of calorimeter hits and tracks and then merged into the hit and track collections of the original event. The particle flow and high level reconstruction is then performed on the merged event.

3. **Particle Flow Level (*Particle Flow* Embedding)**
   The highest level where the merging can be performed is the level of reconstructed particles. With the event reconstruction in the simulated event run up to the point of particle reconstruction, the particle collections of original and simulated event are merged.

For all methods, the simulated event only contains the embedded particles. Otherwise, the detector is empty. This can introduce biases in the event reconstruction, since some parts of the reconstruction are sensitive to noise and pileup in the event. Therefore, the embedding should take place on the lowest possible level of event reconstruction so the largest possible part of event reconstruction is performed on the embedded event. On the other hand, the earlier the merging is performed, the more objects need to be merged and therefore the merging can become more challenging.

The embedding procedure is implemented in CMSSW with two different embedding algorithms. These are called *Particle Flow* embedding (PF) and *Rec-Hit* embedding (RH), based on the above described level of merging of simulated $Z \to \tau\tau$ and original $Z \to \mu\mu$ event. Figure 3.2 shows the scheme of both embedding methods and how the embedding procedure is implemented accordingly.

For reasons of the selection of input data as described in Chapter 3.3, both methods start with a $Z \to \mu\mu$ data event where the full reconstruction including the particle flow algorithm has already been executed. From this event, a separate $Z \to \tau\tau$ event is created where the simulated τ-leptons have the same kinematic properties as the muons they will replace. In both embedding methods the detector simulation and the reconstruction of calorimeter hits and tracks is run on the simulated event.

In case of the PF embedding, the particle flow algorithm is run on the simulated event. In the original $Z \to \mu\mu$ event, the muons and their tracks are removed and the tracks and particles from the simulated $Z \to \tau\tau$ event are merged into the remainder of the original event thus creating a $Z \to \tau\tau$ data/MC hybrid event.

In case of the RH embedding, the hits and tracks of the reconstructed muons are removed and the simulated $Z \to \tau\tau$ event is only reconstructed to this level. The simulated tracks and calorimeter hits are then merged with the remainder of the hits and tracks from the original event. The particle flow objects from the original event are discarded and the particle flow algorithm runs a second time, now on the merged event.

Finally, in both methods, high level objects are reconstructed. This includes the clustering of jets, reconstruction of hadronic taus from jets and calculating the missing transverse energy of the embedded events.

The PF embedding has been studied and commissioned in CMS and was used successfully, for example in the $H \to \tau\tau$ analysis. With increasing pileup the PF embedding is expected to lead to higher reconstruction efficiencies compared to data [24]. Reason for this is that the particle reconstruction of the simulated $Z \to \tau\tau$ event is done in an otherwise empty detector and therefore neglects the effect that pileup has on the particle reconstruction.

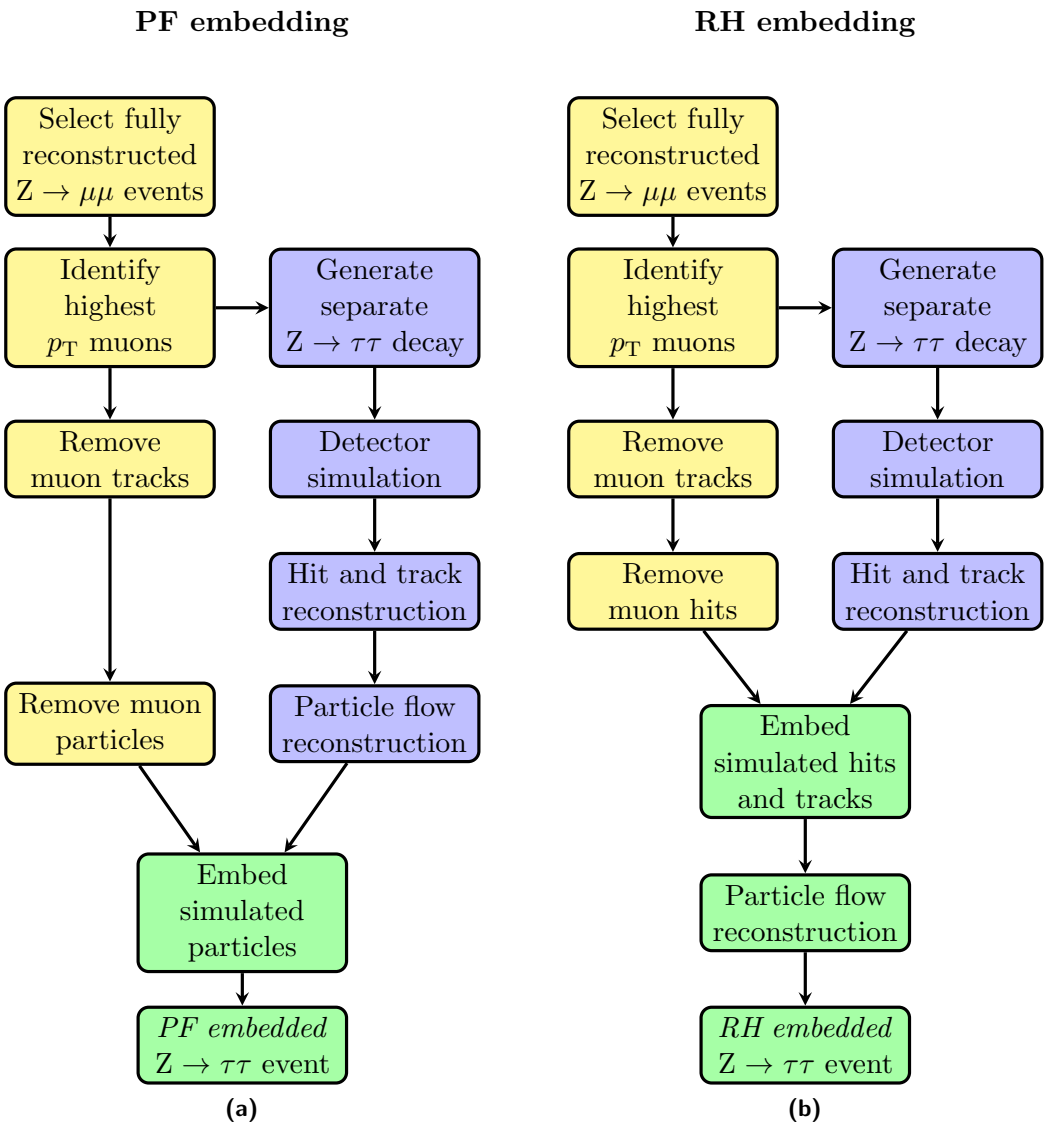


Figure 3.2: Scheme of the implementation of the embedding procedure in the Particle Flow (PF) and Rec-Hit (RH) embedding algorithms. Starting from a selected $Z \to \mu\mu$ event, the highest p_T muons of the event are identified. Both methods then create a separate $Z \to \tau\tau$ decay where the τ-leptons get the same four momentum and initial vertex as the before reconstructed muons. In the PF embedding (3.2a), the muon tracks and particles are removed and the simulated event is reconstructed up to the level of particles. Then the tracks and particle collections of both events are merged, creating a *PF embedded* $Z \to \tau\tau$ event. In the RH embedding (3.2b), the muon tracks and hits are removed and the simulated event is reconstructed up to the level of hits and tracks. Then the tracks and hit collections of both events are merged before the particles are reconstructed in the merged event, creating a *RH embedded* $Z \to \tau\tau$ event.

The RH embedding was introduced trying to compensate for this expected bias, since there the particle flow algorithm is executed on the embedded event. With the hits and tracks of the original event considered, the particle reconstruction will not suffer from effects due to missing pileup. A possible bias in the RH embedding comes from the second execution of the particle flow algorithm that acts on the original hit collections a second time. One disadvantage of the RH embedding is that it needs the larger, so called *RECO* event format since only in this format the hit collections are accessible that are needed for the merging of the simulated event and the original data event.

For the sake of identifying biases inherent to the embedding algorithms, the embedding process can be modified so muons instead of τ-leptons are generated in the simulated event. Thereby, biases introduced by the algorithms themselves as well as biases regarding the reconstruction of muons can be studied separately from biases on the simulation and reconstruction of the τ-leptons.

3.3 Selection of Input Data

The embedding process starts with a preselection of $Z \rightarrow \mu\mu$ events. To retrieve these events from data, a di-muon selection is applied, designed to filter events with two well reconstructed and isolated muons.

The RH embedding needs to modify the calorimeter hit- and the track collections of each selected event. This is only possible in the RECO event format, which contains the reconstructed low level objects needed for the merging and reprocessing of the these collections. The disadvantage of this event format is its event size. A single event with on average 25 pileup vertices will be of the size of ≈ 2 MB , which is about five times as much as the more commonly used pruned *AOD* event format. The large event size of the RECO event format makes an efficient di-μ preselection of $Z \rightarrow \mu\mu$ events desirable. This way computing resources, especially storage space, can be saved.

To pass the selection, muons have to fulfil among others the criteria of the so called *tight* muon selection[18]. This selection includes a number of requirements to reduce the rate of particles that were falsely reconstructed as a muon and to ensure that important quantities, like the transverse momentum, were well reconstructed. This includes for example a minimum requirement on both the number of matched muon stations in the muon chambers and the number of pixel hits in the inner track detector. This way, the number of hadronic particles that punch through the HCAL and are misidentified as muons is reduced, because they are highly unlikely to have a pixel hit in the inner track detector and at the same time punch through more than one layer of absorber in the muon chambers. By requiring at least 5 layers of the inner tracker with hits it is ensured that the uncertainty in reconstructing the curvature and therefore the transverse momentum of the muons is small.

Additionally to the tight muon selection, a selection on the relative charged hadron isolation $R^{\mu,ch}$ is performed where the isolation component $I^{\mu,ch}$ therein was

calculated only based on the charged hadron component of the lepton isolation I^l as given in Equation (2.7). The charged hadrons therein are only the ones that were identified to originate from the primary vertex. The charged hadrons associated with pileup are accumulated in a separate particle collection. When not explicitly stated, the *charged hadrons* will always refer to the ones from the primary vertex. This selection also reduces the rate of misidentified particles, especially from hadronic interactions. The additional requirements on transverse momentum, pseudorapidity and the HLT requirement are introduced to select events that are of interest for further analysis. By requiring an invariant di-muon mass larger than 20 GeV, muons that originate from meson decays are suppressed. The full set of selection criteria and their values is listed in Table 3.1.

Since the embedding also has to be validated without data for example to test software updates in CMSSW and the necessary adaptations in the embedding algorithms, it was designed to be able to process generated MC events as well. This way, it can be tested and validated independently of the availability of data events. The process of generating these MC events is described in Section A of the appendix. The di-muon selection is applied to MC events in the same way.

3.4 Muon Momentum Vector Transformation

The requirement on the relative isolation from charged hadrons, $R^{\mu,ch}$, in the di-muon selection biases this particle collection towards lower numbers of charged hadrons. This can be seen in Figure 3.3a, which shows the average number of charged hadrons around muons in hollow $\eta - \phi$ cones. The distance between muon and charged hadron is calculated from the distance $\Delta R_{\mu,\mathrm{ch}} = \sqrt{(\Delta\eta_{\mu,\mathrm{ch}})^2 + (\Delta\phi_{\mu,\mathrm{ch}})^2}$ in the η- and ϕ-plane. The bin width of 0.02 corresponds to hollow cones of thickness $\Delta R_2 - \Delta R_1 = 0.02$.

The average number of charged hadrons is illustrated for the PF and RH embedding methods as well as for an unbiased MC simulation as point of reference. With increasing values of ΔR, the cones cover larger volumes and therefore more charged hadrons will be contained in the hollow cones. This leads to the approximately linear increase of the average number of charged hadrons with increasing values of $\Delta R_{\mu/\mathrm{ch}}$ in the unbiased MC dataset.

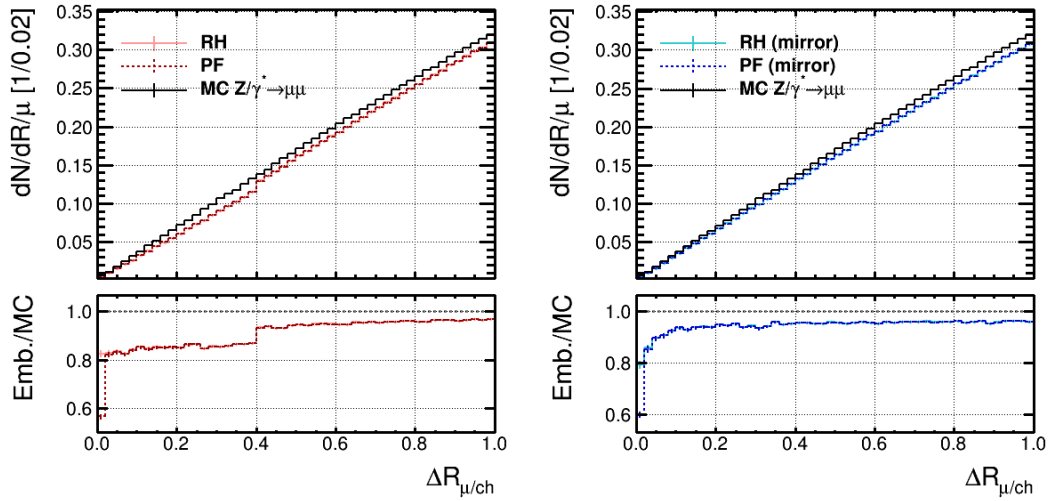
The discrepancies between the MC validation dataset and the embedding methods can be seen best in the ratio plot. This shows the ratio of the average densities of charged hadrons around the muons in the embedded datasets compared to the plain MC simulation.

The embedding methods have an overall reduced number of charged hadrons in the events, especially up to $\Delta R_{\mu/\mathrm{ch}} = 0.4$. The decreased number of charged hadrons up to this radius is mostly caused by the requirement on $R^{\mu,ch}$ in the di-muon selection, since for the calculation of the isolation component $I^{\mu,ch}$ therein, all charged hadrons within the cone $\Delta R_{\mu/\mathrm{ch}} < 0.4$ are considered. Therefore, the area where the τ-leptons

Single muon selection criteria		
Criterion		Requirement
Tight Muon Selection	Global Muon	True
	Particle Flow Muon	True
	$\frac{\chi^2}{ndof}$ of global muon track	< 10
	Number of matched muon stations	> 1
	Number of pixel hits	> 0
	Number of tracker layers with hits	> 5
	Transverse impact parameter d_{xy} of tracker	0.2 cm
	Longitudinal impact parameter d_z of tracker	0.5 cm
	Number of muon chamber hits included in the global muon track fit	> 0
Relative charged hadron isolation $R^{\mu,ch}$		< 0.1
Transverse momentum p_{T}		$> 8\,\mathrm{GeV}$
Pseudorapidity η		< 2.5
Primary vertex category		good primary vertex
Double muon selection criteria		
Criterion		Requirement
HLT requirement		$\mu(17)$ & $\mu(8)$
p_{T} of the highest energetic muon		$> 17\,\mathrm{GeV}$
Invariant di-muon mass		$> 20\,\mathrm{GeV}$

Table 3.1: Requirements for the di-muon selection of input data for the embedding procedure. These are divided into requirements on single muons and requirements on the di-muon system. The selection on χ^2 over the number of degrees of freedom, $ndof$, the number of matched muon stations and the number of pixel hits in the inner tracker filter out a large fraction of hadronic particles that were misidentified as muons. The requirements on the impact parameters d_{xy} and d_z ensure that most of the reconstructed muons originate directly from the primary vertex. As isolation requirement, a selection on the relative charged hadron isolation $R^{\mu,ch}$ is applied. This selection also reduces the rate of misidentified muons. The selection only on the charged hadron component prevents biases for later applied analyses, where the exact composition of the neutral hadron, photon and pileup components to the isolation I^l might be different. The requirement on the transverse momentum, p_{T}, and the HLT requirement are mainly to filter out events that cannot be analysed since they are too low energetic and poorly distinguishable from pileup and other decay products in the events. The requirement on the pseudorapidity, η, is to filter out events that are outside of the most sensitive regions of the detector. The requirement on the invariant mass of the di-muon system is introduced to filter out events that are unlikely to originate from a Z boson decay.

are embedded in the original event are on average more isolated. As a consequence, this will lead to too high selection efficiencies when using the embedding.



(a) (b)

Figure 3.3: Figure 3.3a illustrates the average number of charged hadrons around muons in cones of thickness $\Delta R = 0.02$. The MC dataset as point of reference is shown in black, the RH embedding in bright red and the PF embedding as dotted dark red line. The ratios of the number of charged hadrons of the embedding methods and the MC simulation are shown in the subplot. The MC validation dataset shows an approximately linear increase of the average number of charged hadrons with increasing $\Delta R_{\mu/\mathrm{ch}}$. In comparison, the embedding methods show an overall reduced number of charged hadrons, especially up to $\Delta R_{\mu/\mathrm{ch}} = 0.4$. The step at $\Delta R_{\mu/\mathrm{ch}} = 0.4$ is caused by the requirement on $R^{\mu,ch}$ in the di-muon selection. Apart from entries in the first bin, both embedding methods show the same distribution of charged hadrons. Figure 3.3b shows the density of charged hadrons around muons after applying the mirroring. The mirrored RH embedding is shown in bright blue, the mirrored PF embedding as dotted, dark blue line. The step in the density at $\Delta R_{\mu/\mathrm{ch}} = 0.4$ is removed, thereby improving the overall agreement between the embedding methods and the unbiased MC simulation. Apart from the differences close to the muons, also here both embedding methods show the same charged hadron density.

One possibility to reduce this bias from the selection on $R^{\mu,ch}$ is to mirror the momentum vectors of the muons on the plane defined by the Z boson momentum axis and the beam of colliding protons. In the laboratory frame, this corresponds to swapping the muons in the $r - \phi$ plane with respect to the Z boson momentum axis. This transformation is illustrated in Figure 3.4.

Due to the mirroring, the muons point in a different direction of the detector that was not considered explicitly in the di-muon selection. Therefore, this new area is less biased. Figure 3.3b shows the average number of charged hadrons around the muons

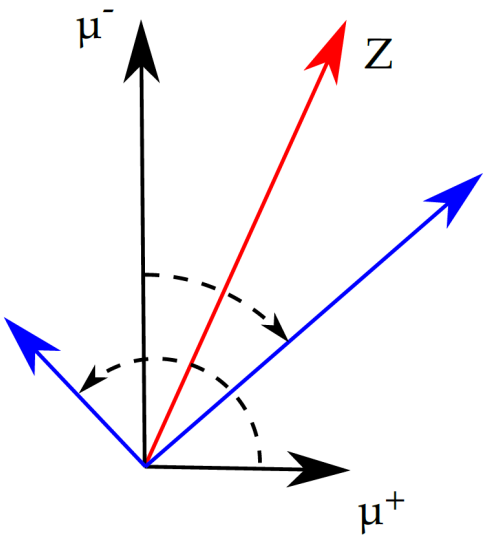


Figure 3.4: Illustration of the mirroring of muons around the Z boson momentum axis [24]. With the mirroring, the new muon momentum vectors are obtained. These point in a new direction of the detector that is less biased from the di-muon selection.

after the mirroring. Compared to the embedding without applying the mirroring transformation, the number of charged hadrons up to $\Delta R = 0.4$ is increased so only a reduced, mostly constant bias remains.

The still visible lower number of charged hadrons in the first bins of the PF and RH embedding is not caused by the isolation requirement. A possible origin are events where the muons emit final state radiation. In these events, a fraction of the final state radiation photons will convert into leptons that can then be reconstructed as charged hadrons. In the embedding without mirroring, these tracks and the corresponding charged hadrons are removed when the muon signal is removed from the original event. Since photons are reconstructed based on energy deposits, they do not have any track information. Therefore, photons cannot be definitively distinguished to originate from noise, pileup or the primary collision vertex [25]. As a result, when mirroring the muons, the final state radiation located around the muon cannot be picked up and mirrored. Therefore, the reconstructed charged hadrons originating from final state radiation would be missing in the mirrored embedding methods as well.

In the RH embedding, additional charged hadrons are reconstructed when the particle flow algorithm processes the merged event. The precise origin of this effect is unknown, but a possible explanation is the linking of tracks from the original event to energy deposits from the embedded muon in the calorimeter. This way, new charged hadrons close to the muon would be reconstructed. This could also explain the lower number of charged hadrons in the PF embedding, where no additional charged hadrons would be reconstructed since the particle flow algorithm processes only the simulated event.

When transforming the muon momentum vectors, it has to be ensured that the underlying Z boson decay remains invariant under the transformation. This invariance has been confirmed using leading order MC simulation with the event generators pythia and madgraph [24].

The effects of the mirroring for both embedding algorithms will be studied in detail in the following chapter.

4 Validation with Muons

A variation of the embedding process is the embedding of muons instead of τ-leptons. This so called *muon* embedding takes the reconstructed muons of the original event and replaces them with muons of the same kinematic properties. This allows for the study of biases inherent to the embedding algorithms and the reconstruction of muons, separately from effects related to the simulation and reconstruction of τ-leptons.

The PF and RH embedding algorithms were originally implemented in CMSSW 5.3. In preparation for the second data taking period at the LHC, various adjustments and optimisation were introduced into the event reconstruction algorithms of CMSSW, for example to cope with the higher instant luminosity and an increased rate of pileup. Additionally from CMSSW 7.0 on, the particle flow algorithm was integrated more tightly with the reconstruction of electrons and photons. This required an updated treatment in the PF embedding algorithm. To make sure the necessary adjustments to the embedding algorithms were successful, the embedding in CMSSW 7.0 is validated using the muon embedding.

The events used for the validation of the muon embedding are taken from a dataset from the MC simulation of the Drell-Yan process with two muons in the final state. The dataset was generated with the matrix element generator *madgraph* [26] and the event generator *pythia 8* [27]. Starting from the event on MC generator level, the generated Drell-Yan events were mixed with pileup events. The detector response was then simulated and reconstructed to the level of the RECO event format as described in Section A of the appendix. The validation dataset was taken from this processed dataset without any further preselection applied. The input data for the different embedding algorithms was obtained by applying the previously introduced di-muon selection on this dataset.

By choosing a subset of the validation dataset as input for the embedding, biases inherent to the embedding methods will be better visible since statistical fluctuations will be partially correlated. In the PF and RH muon embedding, the same input events were used. Thus, statistical fluctuations between the embedding methods will be fully correlated. This will help identifying individual differences between the algorithms.

The baseline event selection that was set up to evaluate the muon embedding is introduced in Section 4.1. The muon embedding is then evaluated in Section 4.2 with this baseline event selection. Eventually, the distribution of the kinematic properties and the reconstructed di-muon mass of the embedded datasets and the validation dataset are compared in Chapter 4.3.

4.1 Baseline Event Selection

For the validation of the embedding algorithms in the new software environment, a baseline event selection was set up. To reduce biases from the di-muon selection of events for the embedding, the analysis uses the same or tighter selection criteria as this preselection.

The selection parameters for the pseudorapidity, η, the transverse impact parameter, d_{xy}, and the longitudinal impact parameter, d_z, were taken from the $H \to \tau\tau$ analysis, introduced in Chapter 2.4. The $\Delta\beta$-corrected relative muon isolation, R^{μ}, was required to be smaller than 0.1, similar to the value used for electrons and muons in most of the six main di-τ decay channels of the $H \to \tau\tau$ analysis.

A new criterion in the baseline selection, compared to the di-muon selection, is the generator matching of muons. The generator matching checks if a muon from MC event generation is within an $\eta - \phi$ cone of $\Delta R = 0.005$ around the reconstructed muons. The reconstructed muons are discarded if no generator level muon was found within this cone. This way, reconstructed muons that do not arise from the Z boson decay are suppressed. Thereby a systematic uncertainty is reduced and inaccuracies of the embedding algorithms will become more significant.

In the di-muon selection, only events with an invariant di-muon mass larger than 20 GeV were selected. The used MC dataset only contains events with an invariant di-muon mass larger than 50 GeV. Therefore, the analysis can be considered to have a requirement of 50 GeV on the invariant di-muon mass and the requirement from the di-muon selection did not need to be introduced. The full list of the applied baseline selection criteria and selection values is given in Table 4.1.

4.2 Baseline Selection Efficiency

The purpose of the embedding is the reduction of systematic uncertainties in the background estimation compared to a purely MC driven background estimation method. At the same time, the embedding procedure introduces new systematic biases, e.g. from inaccuracies when selecting and removing the tracks from the muons of the original event. Therefore, systematic biases of the embedding methods need to be studied first in MC simulation. One example for such a systematic bias are deviations in the event selection efficiency.

The selection efficiency of embedded datasets compared to the selection efficiency of an unbiased dataset is studied in this chapter. To quantify the selection efficiency, a common basis of comparison for the different datasets is established. This basis are the events that remain after the application of the HLT requirement, the requirements on the transverse momentum and the pseudorapidity as well as the generator muon matching as introduced in the baseline event selection. The HLT requirement is included, since it affects distributions like the pseudorapidity and the transverse momentum of the muons as well as the pileup distribution of the events. Therefore, it has to be considered in the validation dataset as well. The additional kinematic

Baseline Selection Criteria		
Criterion		Requirement
Preselection	HLT requirement	$\mu(17)$ & $\mu(8)$
	Transverse momentum p_{T_μ}	$> 20\,\mathrm{GeV}$
	Pseudorapidity η_μ	< 2.1
	Generator muon matching radius ΔR	0.005
Transverse impact parameter d_{xy_μ}		$< 0.045\,\mathrm{cm}$
Longitudinal impact parameter d_{z_μ}		$< 0.2\,\mathrm{cm}$
Muon Selection		Tight
$\Delta\beta$-corrected relative isolation R^μ		< 0.1
Number of muons in event after previous selections applied		2

Table 4.1: Requirements for the baseline event selection. A generator muon matching is introduced to prevent additional systematic uncertainties from particles misidentified as muons. The HLT requirement is taken from the di-muon selection. The transverse momentum, p_{T}, is required to be at least 20 GeV for each muon, tighter than in the di-muon selection and similar to the requirements for leptons in the most significant decay channels of the $H \to \tau\tau$ analysis. To pass the analysis, exactly two muons in an event must have passed the selection criteria. The first four selection criteria are applied to establish a common basis for the embedding methods and the MC validation dataset.

requirements are applied to remove biases from the corresponding requirements in the di-muon selection. This basis of events will be referred to as *preselection* in the following.

Another important quantity in the validation of the embedding is the number of reconstructed pileup vertices, nPU. Pileup dependent deviations of the selection efficiency would also introduce new systematic biases that needed to be accounted for by the introduction of new systematic uncertainties. Especially in the PF embedding, the particle flow reconstruction is expected to perform better in the simulated events without pileup.

Figure 4.1a shows the event selection efficiency over the number of reconstructed pileup vertices of the PF embedding algorithm and the MC validation dataset. The selection efficiency is calculated by dividing the number of events after the full baseline selection by the number of events in the preselection of each individual dataset.

The event selection efficiency is then evaluated by comparing the efficiencies of the embedding and the validation dataset. Their ratio is shown in the ratio plot of Figure 4.1a. The PF embedding shows a flat, approximately 10 % too high reconstruction efficiency from 1 to 61 reconstructed pileup vertices. A significant

increase in the reconstruction efficiency with increasing pileup, as it was expected in the PF embedding algorithm, is not visible in the muon embedding.

The event selection efficiency decreases with increasing pileup up to approximately 10 %. This drop is larger than in previous embedding validations and most likely caused by the out of time pileup mitigation in CMSSW 7.0. This software version still uses out of time pileup mitigation algorithms optimised for bunch crossing rates of 50 ns, while the dataset for the validation was generated under nominal conditions for the second data taking period with collisions that are only 25 ns apart. Thus, the out of time pileup suppression is sub optimal which eventually leads to higher noise in the events and therefore worsened selection efficiencies.

To find the origin of the deviation in the selection efficiency of the PF embedding, different parameters of the baseline event selection were varied. The green line in Figure 4.1a shows the event selection efficiency when only applying the $\Delta\beta$-corrected isolation criterion, R^{μ}, in addition to the criteria for the preselection. Compared to the full selection the relative difference is the order of a few per cent.

Figure 4.1b shows the event selection efficiency with all selection criteria but the $\Delta\beta$-corrected isolation applied. The event selection efficiency is between 96 % and 98 %, with deviations between the embedding and the validation dataset smaller than 2 %.

From the comparison of both plots, it becomes clear that the $\Delta\beta$-corrected isolation is the cause of the drop in the selection efficiency as well as the largest deviations in the selection efficiency compared to the validation dataset. Consequently, the differences in the individual particle collections that contribute of the $\Delta\beta$-corrected muon isolation are studied in the following.

For each particle collection that contributes to the calculation of the $\Delta\beta$-corrected isolation I^{l}:

- charged hadrons from the primary vertex (ch)
- neutral hadrons (nh)
- photons (ph)
- charged hadrons from pileup vertices (ch, PU)

a transverse momentum profile is created. These illustrate the average amount of transverse momentum from the corresponding particles in hollow $\eta - \phi$ cones around the muon.

The transverse momentum density profiles were created based on the events contained in the preselection. In each particle collection, xy, particles up to a cone size of $\Delta R_{\mu/xy} = 0.4$ are considered, corresponding to the cone size used when calculating the individual relative isolation $R^{\mu,xy}$. Thereby, the origin of the discrepancy in the selection efficiency is broken down to the individual components of the muon isolation. The impact of deviations in the p_{T} profiles on the event selection efficiency is estimated by requiring a value of the corresponding isolation criterion $R^{\mu,xy} < 0.1$ additionally to the preselection. Then, the relative event

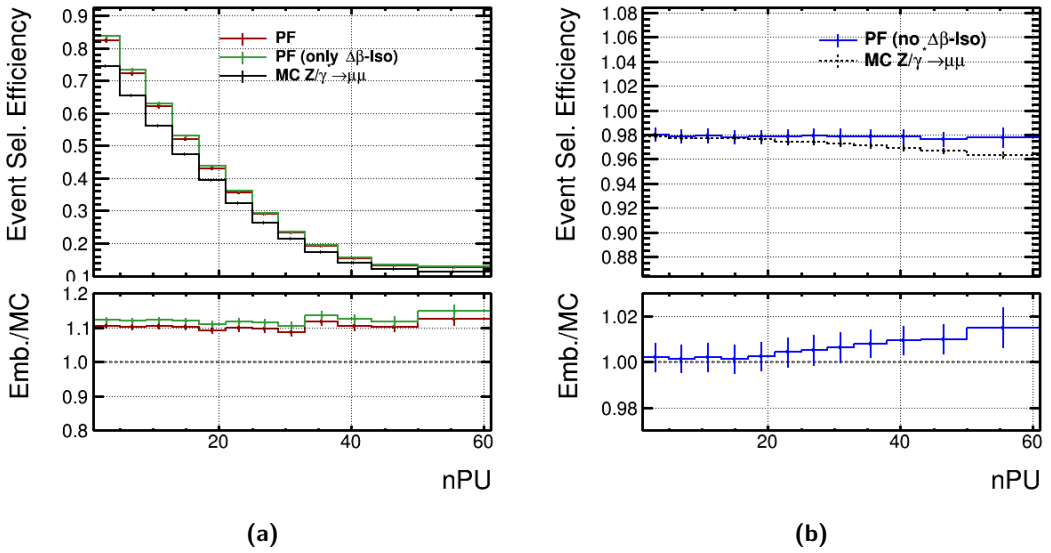


Figure 4.1: Event selection efficiencies over the number of reconstructed pileup vertices, nPU. The red line in Figure 4.1a shows the selection efficiency of the PF embedding when applying the full baseline selection. The green line in this plot shows the selection efficiency when only applying the $\Delta\beta$-corrected isolation in addition to the criteria for the preselection. The validation dataset of $Z/\gamma^* \to \mu\mu$ events is shown in black. The efficiency of these two selections on the PF embedding dataset compared to the efficiency of the validation dataset is approximately 10 % higher. Figure 4.1b shows the corresponding event selection efficiency when applying all criteria from the baseline selection except for the $\Delta\beta$-corrected isolation. The deviations in the event selection efficiency of these criteria is smaller than 2 %. As a consequence, the largest contribution to the discrepancy between the MC validation dataset and the PF embedded dataset in Figure 4.1a is caused by the $\Delta\beta$-corrected isolation.

selection efficiencies of the different embedding methods and the validation sample are compared.

In all particle collections, the PF and RH embedding algorithms will be evaluated with and without the mirroring of the muons, leading to a total of four different embedding methods that are compared to the validation dataset. The mirrored embedding methods will be plotted in shades of blue. The unmirrored embedding methods in shades of red. The RH embedding is drawn as continuous line, whereas the PF embedding is drawn as dotted line. The MC $Z/\gamma^* \to \mu\mu$ validation dataset is drawn in black. The ratio of the distributions of each embedding method and the MC dataset will be shown in a ratio plot in the colours of the corresponding embedding method.

Charged Hadrons

The transverse momentum profile of charged hadrons from the primary vertex is given in Figure 4.2a. All embedding methods show a lower amount of transverse momentum from charged hadrons around the muons, especially the unmirrored embedding methods. The mirroring reduces the bias in the transverse momentum density profile of charged hadrons by approximately 20 %.

The event selection efficiency when only requiring a relative charged hadron isolation of $R^{\mu,ch} < 0.1$ additionally to the selection criteria of the preselection is illustrated in Figure 4.2b. The reduced amount of transverse momentum leads to higher selection efficiencies of the embedded datasets compared to the validation dataset. The unmirrored embedding methods have an approximately 10 % higher selection efficiency than the validation dataset, independent of the embedding algorithm. The mirrored embedding methods only show an approximately 2 % higher selection efficiency. Thus, the mirroring reduces the bias in the event selection from the charged hadron component of the isolation by approximately 8 %.

A possible origin of the remaining deviation are events where the mirrored muons point into the direction of one of the muons before the mirroring. In these cases, the muons would again point into a preselected, on average too clean area of the detector.

A contribution from charged hadrons in direct proximity to the muons is missing in the PF embedding. Nevertheless, this difference does not have a significant impact on the event selection efficiency, where PF embedding and RH embedding do not show any significant differences. Therefore, this deviation in the transverse momentum profile was not studied any further.

Neutral Hadrons

The transverse momentum profile of neutral hadrons is given in Figure 4.3a. The unmirrored PF embedding and the mirrored RH embedding show a good agreement with the validation dataset. These three datasets have a local maximum of the transverse momentum around $\Delta R_{\mu/nh} = 0.05$. The origin of this is the particle flow algorithm. In the presence of a muon, the algorithm regroups neutral hadrons close to the muons and adds the expected amount of energy from muon ionisation in the HCAL to the neutral hadrons. This correction creates this additional energy close to the muon. Due to the regrouping of neutral hadrons, this effect gets larger with more neutral hadrons and therefore more pileup in the event.

Since the embedding uses a fully reconstructed dataset as input, this correction is already applied on the level of input data in the selected $\mathrm{Z}/\gamma^* \to \mu\mu$ events. In the case of the RH embedding, this correction is applied a second time when the particle flow algorithm processes the merged event. Therefore, in the unmirrored RH embedding, this correction is applied twice in the same place. This leads to the observable excess of transverse momentum within $\Delta R_{\mu/nh} \lesssim 0.1$. In the mirrored RH embedding, this correction is also applied a second time, but due to the mirroring in

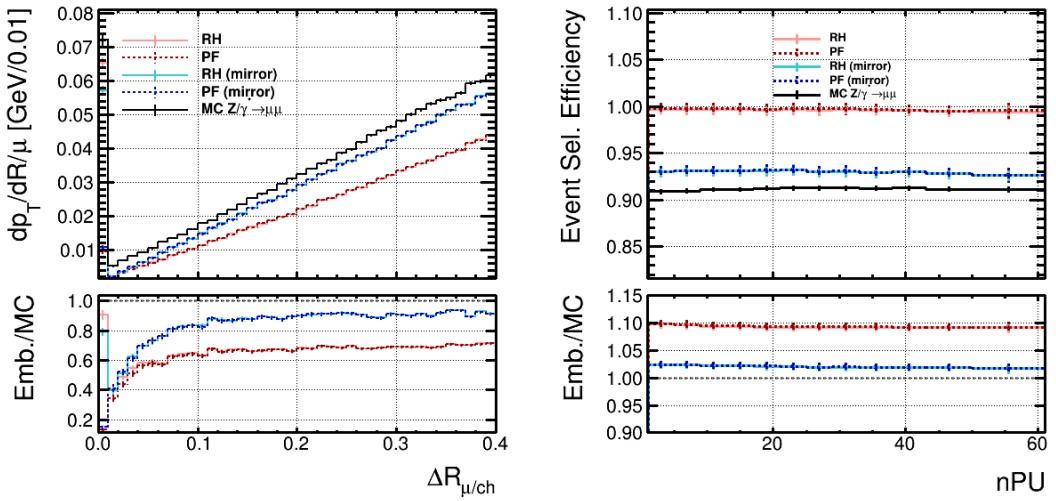
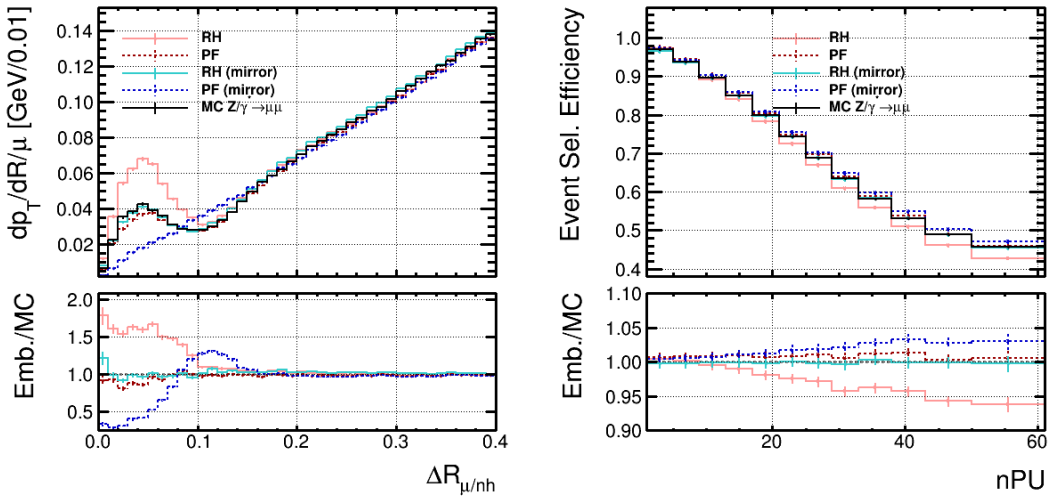


Figure 4.2: Figure 4.2a shows the average amount of transverse momentum of charged hadrons from the primary vertex around muons. The unmirrored embedding methods show a significantly reduced amount of transverse momentum due to the requirement of $R^{\mu,ch} < 0.1$ in the di-muon selection. This difference is reduced by approximately 20 % for $\Delta R > 0.1$ in the mirrored embedding methods. Figure 4.2b illustrates the effect of these deviations on the event selection efficiency. The unmirrored embedding methods show a 10 % too high selection efficiency, independent of the amount of pileup in the event. The mirroring reduces this difference to 2 %. The amount of transverse momentum close to the muons differs in both embedding methods. Nevertheless, this difference has no significant impact on the selection efficiency.

a different area of the detector. Thus, in the RH embedding, the mirroring reduces the deviation in the neutral hadron isolation.

In case of the PF embedding, the particle flow algorithm processes the simulated event. These events are free of pileup and no substantial amount of neutral hadrons is contained in them. Therefore, the correction is not noticeable in the PF embedding. As a result, the effect of the correction is lost in the mirrored PF embedding and the amount of transverse momentum from neutral hadrons in this dataset increases approximately linearly with increasing distance $\Delta R_{\mu/nh}$.

The impact of these deviations on the event selection is illustrated in Figure 4.3b. The plot shows the event selection efficiency when only applying a selection on the relative neural hadron isolation, $R^{\mu,nh}$, additionally to the preselection. The mirrored PF embedding shows an increasing event selection efficiency up to approximately 3 % for large nPU. The unmirrored RH embedding shows an up to 6 % lower event selection efficiency due to the additional transverse momentum close to the muon.



(a) (b)

Figure 4.3: Figure 4.3a shows the average amount of transverse momentum or neutral hadrons around muons. The mirrored RH embedding and unmirrored PF embedding show a good agreement with the validation dataset. Thus, these embedding methods do not show significant deviations in the event selection efficiency, as illustrated in Figure 4.3b. The additional transverse momentum in the unmirrored RH embedding leads to a pileup dependent drop of the selection efficiency up to 6 % around 50 pileup vertices. The mirrored PF embedding shows a pileup dependent increase of the selection efficiency up to 3 % around 50 pileup vertices.

Photons

The transverse momentum profile of photons is given in Figure 4.4a. Here, the unmirrored embedding methods show a better agreement with the MC validation dataset than the mirrored embedding methods. As mentioned before, photons cannot be linked to a specific particle and therefore, the photons from muon final state radiation cannot be picked up and mirrored as well. This leads to a lack of energy close to the muon in the mirrored embedding methods. The largest deviations are within a cone of $\Delta R_{\mu/\mathrm{ph}} < 0.05$.

The impact on the event selection efficiency when only requiring the relative photon isolation $R^{\mu,ph}$ to be smaller than 0.1 is illustrated in Figure 4.4b. The unmirrored embedding methods show a by $2 - 5\,\%$ increased selection efficiency, independent of the amount of pileup in the event. The deviation in the mirrored embedding algorithms increases with increasing pileup from 4 % up to approximately 10 %.

Trying to find a possibility to reduce the impact of the missing final state radiation in the mirrored embedding methods, the correlation between the muon final state radiation and the energy of reconstructed photons was studied. Figure 4.5 shows the amount of muon final state radiation from the MC simulation over the sum of

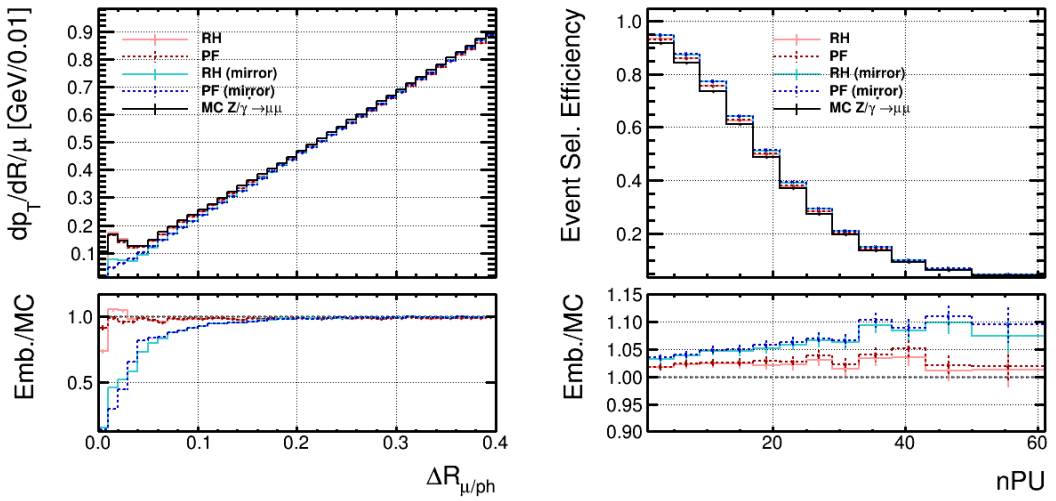


Figure 4.4: Figure 4.4a shows the average amount of transverse momentum from photons around muons. The local maximum around $\Delta R_{\mu/\mathrm{ph}} \approx 0.01$ in the transverse momentum of the unmirrored embedding methods and the MC validation dataset is caused by the muon final state radiation. This contribution is lost when mirroring, since photons have no track information and cannot be linked to a specific particle. Figure 4.4b illustrates the effect of the lost muon final state radiation on the event selection efficiency. The mirrored embedding methods show an up to 10 % increased selection efficiency due to the lack of final state radiation in some events.

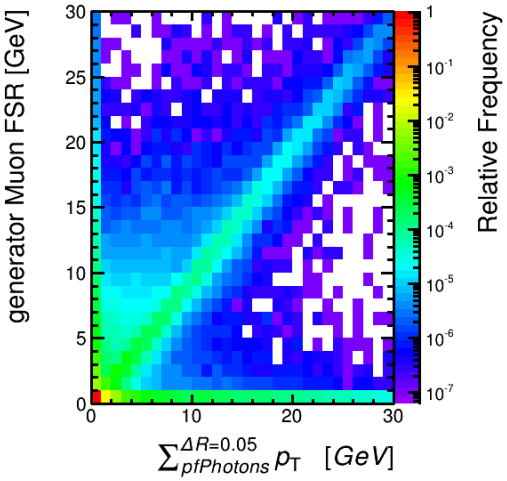


Figure 4.5: Amount of muon final state radiation from the MC event generation as a function of the sum of transverse momentum from reconstructed photons within a cone of $\Delta R_{\mu/\mathrm{ph}} = 0.05$. The first column of bins corresponds to events where the generated muon final state radiation was not reconstructed as a photon within the chosen cone. In events where the muon final state radiation was not reconstructed, no energy is lost in the mirroring. Therefore, these events do not introduce a bias when mirroring. In most cases when the muon final state radiation is reconstructed, the amount of energy within $\Delta R_{\mu/\mathrm{ph}} = 0.05$ correlates with the muon final state radiation.

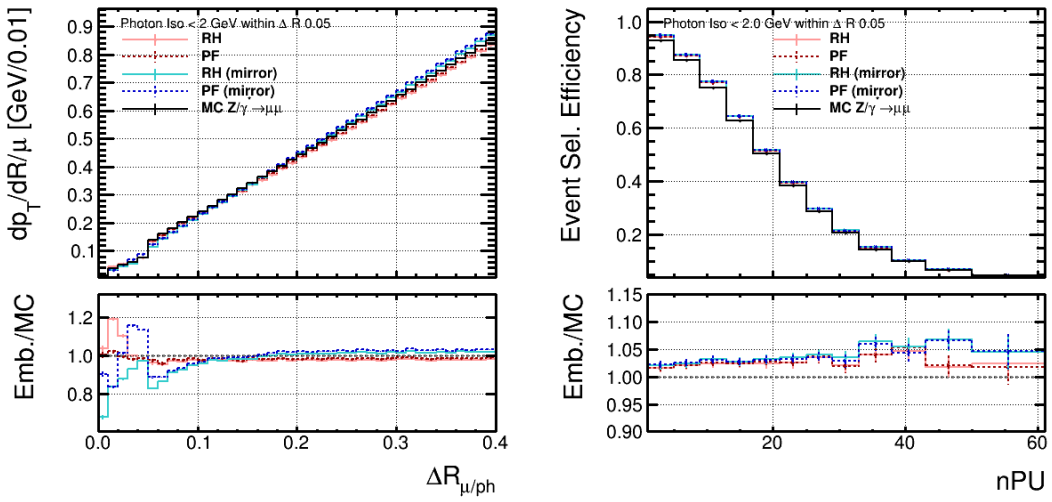
transverse momentum from the reconstructed photons within a cone of $\Delta R_{\mu/\mathrm{ph}} = 0.05$.

If the generated muon final state radiation was not reconstructed as a photon, no muon final state radiation is lost when mirroring. These events are contained in the first column of bins, where less than 1 GeV energy from photons was reconstructed within $\Delta R_{\mu/\mathrm{ph}} < 0.05$. Only where the muon final state radiation was reconstructed as a photon, energy gets lost when mirroring. In these cases, a correlation between the final state radiation and the reconstructed amount of photon energy within the cone of $\Delta R_{\mu/\mathrm{ph}} < 0.05$ is clearly visible.

Due to the correlation, the impact of the final state radiation can be reduced by introducing a requirement on the sum of energy from the photons. A sensible selection value was determined empirically to be 2 GeV for the sum of energy within a cone size of $\Delta R_{\mu/\mathrm{ph}} = 0.05$. This selection removes less than 3 % of the total number of selected events. The impact of this requirement on the transverse momentum profile and the event selection efficiency is illustrated in Figure 4.6.

The impact of the chosen isolation cone size is visible as a small step in the transverse momentum distribution, which otherwise increases linearly with increasing $\Delta R_{\mu/\mathrm{ph}}$. With the requirement, the bias from the mirroring on the event selection

efficiency is almost fully removed and all embedding methods show an approximately $2 - 6\,\%$ increased selection efficiency.



(a) (b)

Figure 4.6: Transverse momentum profile and event selection efficiency when requiring less than 2 GeV on the sum of photon isolation within a cone of $\Delta R_{\mu/\mathrm{ph}} = 0.05$. This selection preferably removes events with large amounts reconstructed muon final state radiation. Thereby, the deviation that was introduced by the mirroring in the event selection efficiency is almost completely removed.

Charged Hadrons from Pileup

The transverse momentum profile of charged hadrons from pileup is given in Figure 4.7a. This shows no significant deviations in any of the examined embedding methods. Thus, the event selection efficiency when only applying a selection on the relative isolation on charged hadrons that emerged from pileup does not show any systematic deviations.

Summary of Isolation and Baseline Selection Efficiency

The impact of the deviations in the individual components on the event selection efficiency is summarised in Table 4.2.

The deviations in the individual particle collections are well understood. Due to the construction of the $\Delta\beta$-corrected isolation, as given in equation 2.7, the neutral hadrons and photons only contribute to the isolation I^l if the sum of their isolation contribution is larger than $\frac{1}{2}$ times the contribution of the charged hadrons from pileup. Therefore, the impact of these deviations in the total deviation of the event selection efficiency is reduced.

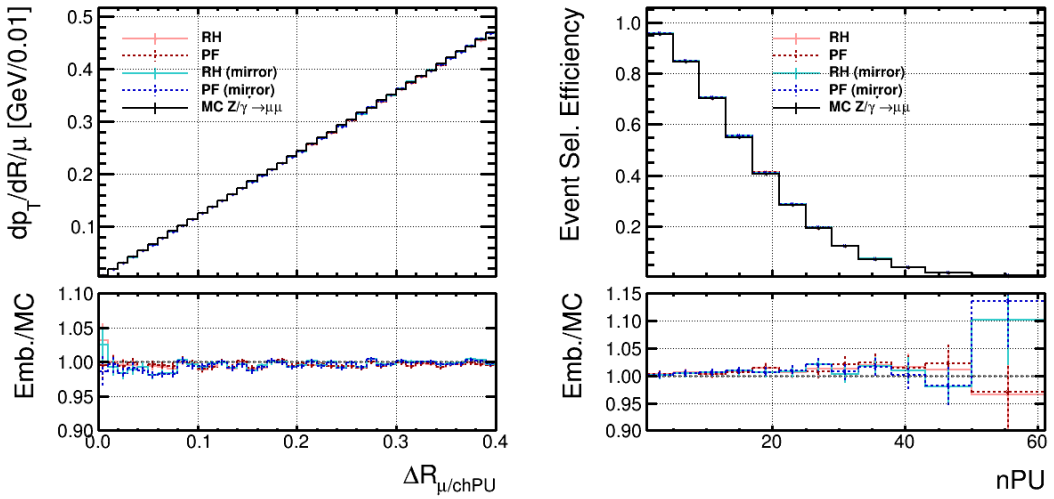


Figure 4.7: The transverse momentum profile of charged hadrons from pileup is shown in Figure 4.7a. No systematic deviations can be found. Thus, the event selection efficiency, shown in Figure 4.7b also does not show any systematic deviations.

The event selection efficiency of the full baseline event selection including requiring less than 2 GeV on the sum of energy from photons within a cone of $\Delta R_{\mu/ph} = 0.05$ is illustrated in Figure 4.8.

The mirrored RH embedding shows the smallest deviations of all studied embedding methods. The event selection efficiency differs between 5 % and 6 % from the selection efficiency of the validation dataset. A clear dependence on the amount of pileup in the events is not evident. The fluctuations in this method are most likely caused by the deviations from the photon isolation component and probably still related to missing muon final state radiation.

The unmirrored PF embedding, shows an approximately constant, 10 % too high selection efficiency, mainly caused by the bias in the charged hadrons from the di-muon selection.

The event selection efficiency of the mirrored PF embedding and the unmirrored RH embedding depends on the amount of pileup in the event. These trends are mainly caused by the biases in the neutral hadron isolation component.

Based on these results, the mirrored RH embedding is the least biased embedding method with the smallest dependence of the event selection efficiency on the number of reconstructed pileup vertices.

Embedding method	Deviations in Selection Efficiency Charged H.	Neutral H.	Photons
PF	+10 %	< +1 %	+2 to +5 %
RH	+10 %	0 to -6 %	+2 to +5 %
Mirrored PF	+2%	0 to +3 %	+2 to +6 %
Mirrored RH	+2%	< +1 %	+2 to +6 %

Table 4.2: Approximate deviations on the event selection efficiency in the individual components of the analysis. The values in the photons are taken from the deviations with applying the criterion to suppress muon final state radiation.

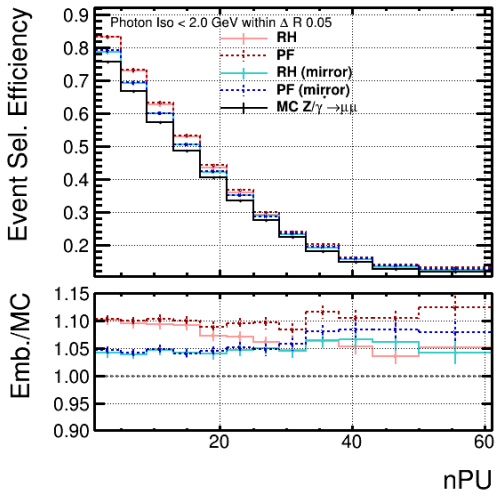


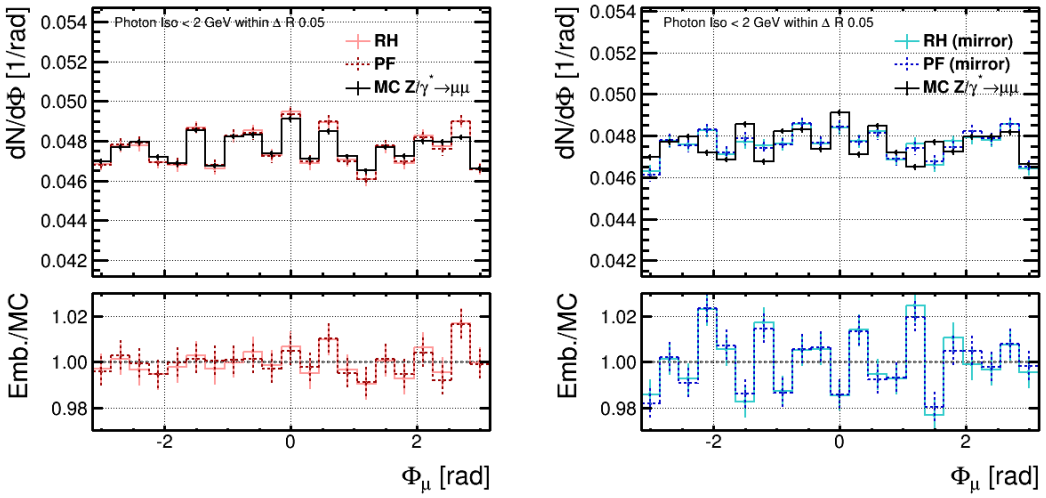
Figure 4.8: Event selection efficiencies of the complete baseline event selection for the studied embedding methods and the validation dataset, including the requirement of less than 2 GeV photon isolation within $\Delta R_{\mu/ph} = 0.05$ to suppress muon final state radiation.

4.3 Kinematic properties

For the evaluation of the embedding, biases on the kinematic properties of the muons need to be considered as well. Figure 4.9 illustrates the distribution of the azimuthal angle, Φ, of the reconstructed muons of the different embedding methods after the baseline event selection. On the left side, Figure 4.9a shows the angular distribution of the PF and RH embedding in comparison to the MC dataset. The distribution of muons in the MC dataset varies by approximately 2 % over the whole range from $-\pi$ to $+\pi$. These fluctuations are caused by small variations in the sensitivity of the detector. The fluctuations of the unmirrored embedding methods are correlated with the fluctuations of the validation dataset. This shows that the unmirrored embedding preserves the Φ distribution. In case of the mirrored embedding methods in Figure 4.9b, the fluctuations are not correlated anymore. This is caused by the mirroring of the muons that changes the direction Φ of the embedded muons. No systematic deviations in the azimuthal angle Φ are observed.

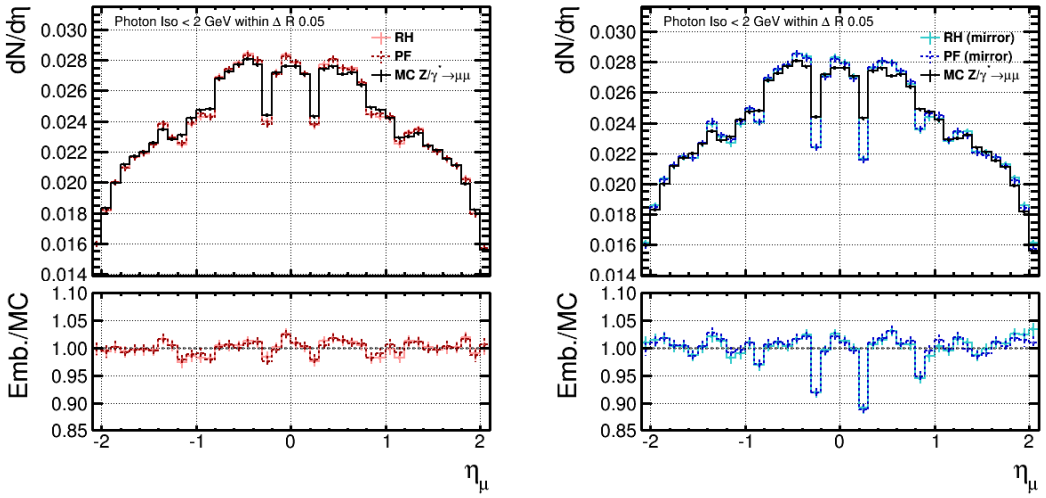
Figure 4.10 shows the distribution of the pseudorapidity, η, for the unmirrored and mirrored embedding methods after the baseline event selection. The MC validation dataset and the unmirrored PF and unmirrored RH embedding datasets are illustrated in Figure 4.10a. The unmirrored embedding shows a good agreement with the validation dataset within the whole selected range of $|\eta| < 2.1$. The drop in the event selection efficiency around $|\eta| = 0.025$ is caused by the service chimneys of the magnet cooling system. One of the service chimneys is located around $\eta = -0.025$, a second chimney around $\eta = +0.025$. Due to these, the muon chambers have a gap and muons within this range of pseudorapidity are therefore less likely to fulfil the requirements for the number of hits in the muon chambers of the tight muon selection. The deviations in the mirrored embedding methods, illustrated in Figure 4.10b, are larger. Due to the mirroring, the azimuthal angle Φ changes, but not the pseudorapidity η. Thereby, some muons within $|\eta| = 0.025$ that did not point into the directions of the gaps in the muon system get rotated into the direction of the gaps. This leads to a further reduced number of events around $|\eta| = 0.025$. Apart from the deviation due to the service chimneys, the overall agreement of the η distributions of all embedding methods is within a few per cent.

The distribution of the transverse momentum, p_{T}, of muons after the baseline event selection is given in Figure 4.11. The plots on the left side show the unmirrored embedding methods, the plots on the right side the mirrored embedding methods. The deviations are within a few per cent with a small trend to lower values of p_{T}. Especially in the mirrored embedding methods, there is a small excess for muons around 40 GeV transverse momentum. This trend is most likely caused by a loss of photons, especially from muon final state radiation, when mirroring. Some of the final state radiation is suppressed by the requirement of 2 GeV on the photon isolation within the cone of $\Delta R = 0.05$. Nevertheless, the loss of a photon of less than 2 GeV can still have a measurable impact on the relative muon isolation, especially for muons of low transverse momentum. Additionally, these muons are more likely to have radiated a final state radiation photon in the first place and are therefore



(a) **(b)**

Figure 4.9: azimuthal distribution of selected muons in the MC validation sample and the unmirrored and mirrored embedding methods. The fluctuations in the unmirrored embedding methods and the validation dataset are partly correlated. This correlation is removed by the mirroring.

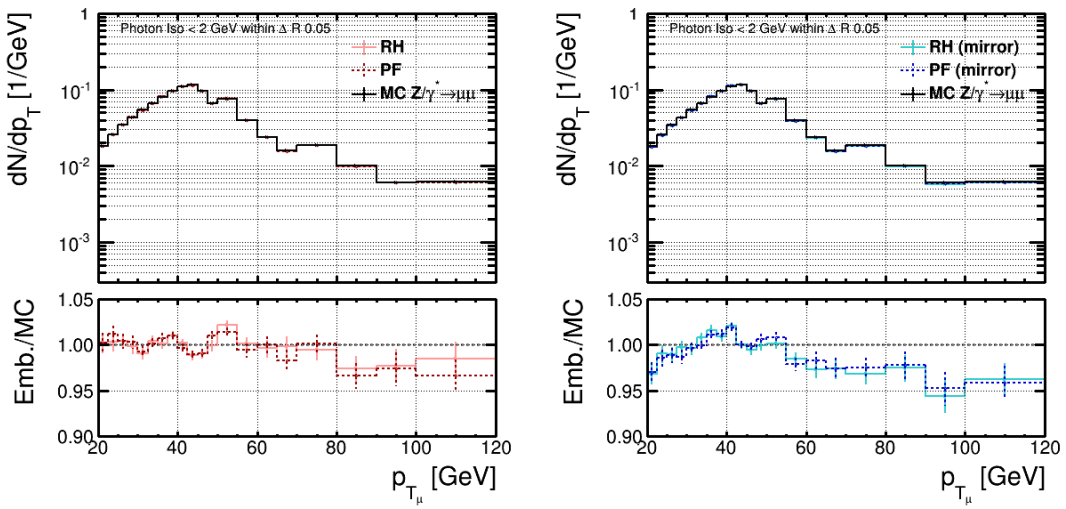


(a) **(b)**

Figure 4.10: Distribution of the pseudorapidity, η, of muons. The muon system has a gap around $\eta = \pm 0.025$ caused by service chimneys for the magnet cooling system. This gap causes a general drop in the selection efficiency in this η region. Since the mirroring changes Φ, but leaves η unchanged, some muons get rotated into the direction of the gap. This causes the additional drop in the mirrored embedding methods for $\eta = \pm 0.025$.

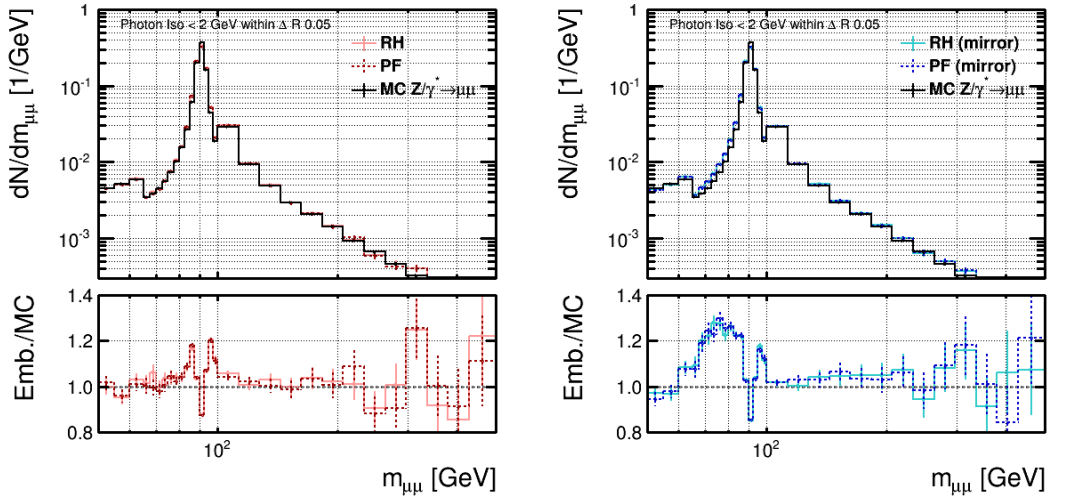
more likely to be biased by the mirroring transformation. The excess in muons with a transverse momentum around 40 GeV is better visible when looking at the reconstructed di-muon mass as illustrated in Figure 4.12.

The peak of the reconstructed di-muon mass is smeared out in the embedded samples, since detector effects are applied twice on the embedded muons. The discrepancies in the mirrored embedding method for masses below the Z boson mass peak is caused by the missing photon from final state radiation when rotating. Apart from these two deviations, the distribution of the reconstructed di-muon mass shows only small discrepancies between the embedding methods and the validation dataset.



(a) **(b)**

Figure 4.11: Distribution of transverse momentum, p_{T} of muons. The deviations between the validation dataset and the embedding methods are within 5 %.



(a) **(b)**

Figure 4.12: Reconstructed di-muon mass of embedded datasets in comparison with a MC validation dataset. The Z boson mass peak is smeared out since detector effects get applied twice on the embedded muons. The mirrored embedding methods additionally show an excess in the number of selected events with di-muon masses between 60 GeV and the Z boson mass peak. Above 100 GeV, all muon embedding methods show a good agreement with the validation dataset.

5 Validation with Tauons

With the muon embedding and occurring deviations well understood, the embedding of τ-leptons is studied in this chapter. This so called *tau* embedding is investigated in the $\mu\tau_\mathrm{h}$ final state where one of the two embedded τ-leptons decays leptonically into a muon and the other τ-lepton decays hadronically. For the PF and RH embedding, different subsets of the $\mathrm{Z} \to \mu\mu$ events were used. Thus, statistical fluctuations between the embedding methods will not be correlated.

The $\mu\tau_\mathrm{h}$ final state was chosen, since it allows to verify effects on the reconstruction of muons that were discovered in the muon embedding and at the same time allows to study deviations related to the simulation and reconstruction of hadronically decaying τ-leptons. Due to the excellent muon reconstruction of the CMS detector, this decay channel is expected to additionally have a higher selection efficiency than e.g. the $e\tau_\mathrm{h}$-decay channel.

The MC dataset used for the validation of the tau embedding consists of $\mathrm{Z} \to \tau\tau$ events that originate from the same dataset as the $\mathrm{Z} \to \mu\mu$ events used for the original events in the embedding process. The $\mathrm{Z} \to \tau\tau$ events were mixed with the pileup from the same dataset and processed using the same reconstruction algorithms that were used for the original $\mathrm{Z} \to \mu\mu$ events of the embedding. This way, discrepancies from different configurations and settings in the event generation and processing are avoided.

The validation dataset for the $\mu\tau_\mathrm{h}$ final state was created, by removing all events in the $\mathrm{Z} \to \tau\tau$ dataset, where the two τ-leptons did not decay into the $\mu\tau_\mathrm{h}$ final state. Additionally, a minimum transverse momentum of 8 GeV for each of the reconstructed τ-leptons was required. This reduced $\mathrm{Z} \to \tau\tau \to \mu\tau_\mathrm{h}$ dataset is taken as the validation dataset for the $\mu\tau_\mathrm{h}$-channel of the tau embedding.

In the embedding, the decay of τ-leptons is simulated with *Tauola*, a MC generator dedicated to generating τ-lepton decays. To increase the yield of the original events, Tauola was configured so only the $\mu\tau_\mathrm{h}$ final state was simulated. Additionally, a minimum transverse momentum of 8 GeV for each of the generated τ-leptons was required.

The ability to set lower boundaries for the transverse momentum of the τ-lepton decay products and the ability to choose a specific final state is one of the strengths of the embedding procedure. This allows to exploit a larger fraction of recorded events compared to a fully data driven background estimation.

Section 5.1 introduces the baseline event selection for the $\mu\tau_\mathrm{h}$-channel. In Section 5.2, the selection efficiency of the baseline event selection of the embedded datasets and the validation dataset are compared, followed by a comparison of the kinematic properties in Section 5.3.

5.1 Baseline Event Selection

Similar to the baseline event selection in the muon embedding, a baseline event selection for the $\mu\tau_{\mathrm{h}}$ final state of the di-τ decay was set up. The muon selection criteria were taken from the baseline event selection of the muon embedding. Additionally, the previously introduced requirement of less than 2 GeV energy from photons within a cone size of $\Delta R_{\mu/\mathrm{ph}} = 0.05$ was required, to suppress the effects of final state radiation.

For the τ-lepton, the requirements on the pseudorapidity, η_τ, the longitudinal impact parameter, d_{z_τ} and the $\Delta\beta$-corrected absolute isolation, I^τ, were taken from corresponding values for the $\mu\tau_{\mathrm{h}}$-channel of the $H \to \tau\tau$ analysis. Only the requirement on the transverse momentum of the reconstructed τ_{h} was loosened to 20 GeV. The reconstructed leptons were matched to be within tight η-ϕ cones around a τ-lepton on MC generator level. This way, misidentifications were suppressed. The muons had to be within cones of $\Delta R < 0.005$, the reconstructed hadronic τ-leptons within $\Delta R < 0.01$.

HLT requirements cannot be used in the tau embedding, since the reprocessing of the high level triggers is not implemented in the embedding algorithms. Therefore, only the HLT requirements of the original $\mathrm{Z} \to \mu\mu$ event would be accessible for the embedded events.

For the calculation of the isolation and the suppression of misidentifications of hadronically decaying τ-leptons, so called *tau discriminators* are calculated in CMSSW. In this analysis the discriminators for decay mode finding, loose electron rejection and tight muon rejection were used.

Hadronic τ-lepton decays are reconstructed with the so called Hadron Plus Strips Algorithm. This algorithm starts from a particle flow jet and checks if at least one charged hadron with hits in the strips of the inner track detector is contained in the jet. Depending on the number of reconstructed hadrons, one of the hadronic τ-lepton decay modes is then assigned to a tau discriminator.

To avoid electronically decaying τ-leptons being misidentified as hadronically decaying τ-leptons, an MVA was trained for the discrimination of electrons and pions. This MVA is evaluated for each jet in the reconstruction of high level objects in CMSSW. When choosing the *loose* electron rejection working point, this MVA needs to yield a value larger 0.6. If this criterion is not fulfilled, the reconstructed τ_{h} is discarded, since it becomes too likely it is a misidentified electronic decay of a τ-lepton.

Similarly to the electrons, the *tight* muon rejection working point was chosen to suppress misidentified muonic τ-lepton decays. For this working point, the leading track used to reconstruct the τ-lepton may not be matched to muon tracks in either the muon chambers or the inner track detector. Additionally, the energy deposit in the ECAL and HCAL must exceed 20 % of the reconstructed energy of the leading track. This way muons are suppressed, since as minimal ionising particles, they are unlikely to deposit more than 20 % of their energy in the calorimeters.

Baseline Selection Criteria		
Muon Selection Criteria		**Requirement**
Preselection	Transverse momentum p_{T_μ}	$> 20\,\mathrm{GeV}$
	Pseudorapidity η_μ	< 2.1
	Generator τ matching radius ΔR_μ	0.005
Transverse impact parameter d_{xy_μ}		$< 0.045\,\mathrm{cm}$
Longitudinal impact parameter d_{z_μ}		$< 0.2\,\mathrm{cm}$
Muon Selection		Tight
$\Delta\beta$-corrected relative isolation R^μ		< 0.1
Photon isolation $I^{\mu,ph}$ within $\Delta R_{\mu/\mathrm{ph}} = 0.05$		$< 2\,\mathrm{GeV}$
Number of muons in event after previous selections applied		1
τ_h Selection Criteria		**Requirement**
Preselection	Transverse momentum p_{T_τ}	$> 20\,\mathrm{GeV}$
	Pseudorapidity η_τ	< 2.4
	Generator τ matching radius ΔR_τ	0.01
Longitudinal impact parameter d_{z_τ}		$< 0.2\,\mathrm{cm}$
τ-lepton discriminators	τ_h decay mode	Assigned
	Electron rejection	Loose
	Muon rejection	Tight
$\Delta\beta$-corrected absolute isolation I^τ		$< 1.5\,\mathrm{GeV}$
Number of τ-leptons in event after previous selections applied		1

Table 5.1: Criteria and requirements for the baseline event selection in the $\mu\tau_\mathrm{h}$-channel. The selection criteria for the muon in the event were taken from the baseline event selection of the muon embedding. Additionally, the new requirement to suppress muon final state radiation is introduced. The selection criteria for the τ-leptons were mostly taken from the values of the $\mu\tau_\mathrm{h}$-channel in the $H \to \tau\tau$ analysis. Only the requirement on the transverse momentum was loosened. A generator matching was introduced, requiring a generator level τ-lepton within a small cone around the reconstructed leptons.

For the isolation, the absolute sum of isolation, I^{τ}, around the reconstructed τ_{h} was required to be smaller than 1.5 GeV. The isolation was calculated using the $\Delta\beta$-corrected isolation with a cone size of $\Delta R = 0.3$. Additionally, the particles contributing to the isolation were required to have at least three hits in the inner track detector. The isolation of a τ_{h} using this set of parameters is calculated and provided by the reconstruction algorithms of CMSSW as a tau discriminator called *RawCombinedIsolationDBSumPtCorr3Hits*.

A summary of the selection criteria is provided in Table 5.1.

5.2 Baseline Selection Efficiency

The four embedding methods are evaluated by comparing the efficiency of the baseline event selection as a function of the number of reconstructed pileup vertices, nPU. As for the muon embedding, a preselection of events builds the basis for evaluating the event selection efficiency. In the tau embedding, the preselection covers the requirements on the transverse momentum, the pseudorapidity and the generator τ-lepton matching. The event selection efficiency when applying the full baseline event selection is given in Figure 5.1. All embedding methods show a between 10 % and 20 % increased selection efficiency compared to the $\mathrm{Z} \to \tau\tau \to \mu\tau_{\mathrm{h}}$ validation dataset. The PF embedding methods show an on average 5 % higher selection efficiency than the RH embedding methods. All embedding methods but the RH embedding without the mirroring show a slight trend towards higher selection efficiencies for increasing amounts of pileup.

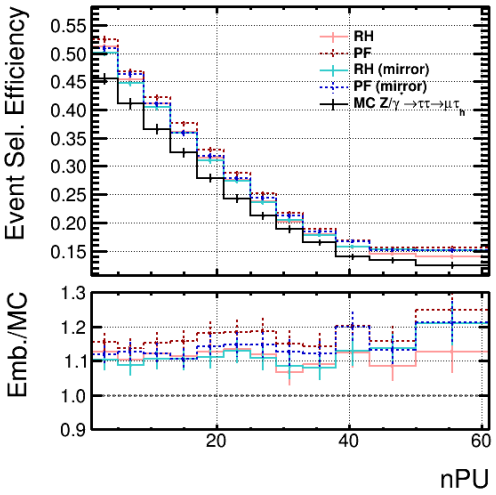


Figure 5.1: Event selection efficiency in the $\mu\tau_{\mathrm{h}}$-channel of the tau embedding. The baseline event selection as introduced in Section 5.1 was applied. All embedding methods show a between 10 % and 20 % increased selection efficiency compared to the $\mathrm{Z} \to \tau\tau \to \mu\tau_{\mathrm{h}}$ validation dataset.

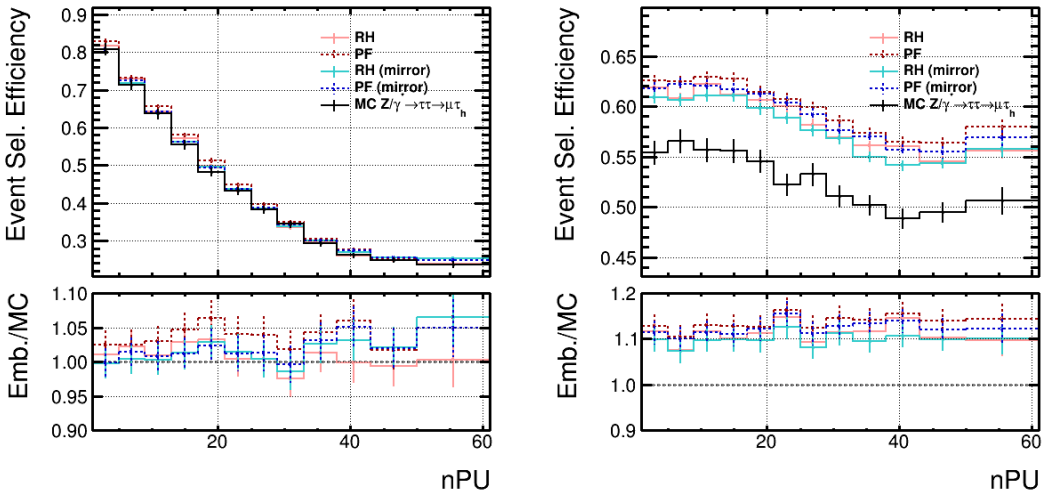


Figure 5.2: Modified event selection efficiencies as a function of the number of reconstructed pileup vertices. Figure 5.2a shows the event selection efficiency when only applying the muon selection criteria as given in Table 5.1. The event selection efficiency of the muons in the embedded datasets is up to 7 % increased compared to the validation dataset. Figure 5.2b shows the event selection efficiency when only applying the τ-lepton selection criteria. All embedding methods show a by 10 % to 15 % increased selection efficiency. The individual deviations from both figures roughly add up to the corresponding total deviation, shown in Figure 5.1.

Looking for the origin of the discrepancies, the baseline selection was modified. Figure 5.2a shows the event selection efficiency when only applying the muon selection criteria. Figure 5.2b shows the event selection efficiency when only applying the τ-lepton selection criteria.

The event selection efficiency of the embedded datasets when only applying the muon selection criteria deviates from the selection efficiency of the validation dataset by up to 7 %. The unmirrored RH embedding shows the best agreement with the validation dataset and no trend towards higher selection efficiencies for larger amounts of pileup in the event. For a more substantial statement, a larger number of events, especially in the validation dataset would be needed.

In the muon embedding, the neutral isolation component introduced a pileup dependent bias in the unmirrored RH embedding that caused a compared to the other embedding methods up to 6 % decreased event selection efficiency with increasing pileup.

Figure 5.3a shows the event selection efficiency when removing the neutral isolation component $I^{\mu,nh}$ from the $\Delta\beta$-corrected isolation. As expected, without the neutral isolation component, the selection efficiency of the unmirrored RH embedding increases compared to the other embedding methods and their selection efficiency

differ less. This demonstrates, how the effects on the muon reconstruction that were discovered in the muon embedding also have a measurable effect in the tau embedding.

Figure 5.3b shows the event selection efficiency for the embedding methods when applying all muon selection criteria but the isolation requirement. Without the isolation requirement, all embedding methods agree better than 2 % with the validation dataset. Therefore, the isolation remains the main contribution to deviations in the event selection efficiency regarding muons.

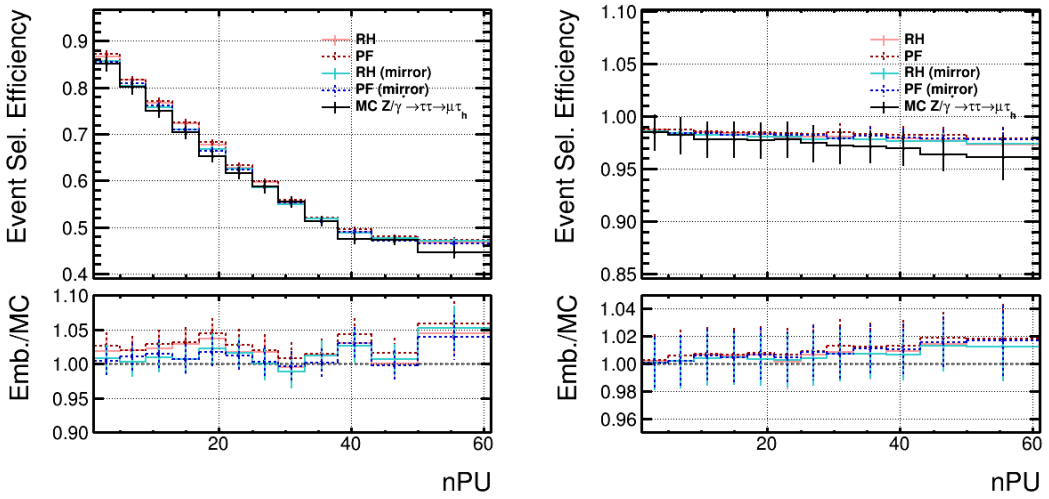


Figure 5.3: Modified event selection efficiencies as a function of the number of reconstructed pileup vertices. Figure 5.3a shows the event selection efficiency when only applying the muon selection criteria and additionally removing the contribution from the neutral isolation component $I^{\mu,nh}$ from the $\Delta\beta$-corrected isolation. The in Figure 5.2a visible, allegedly better agreement between validation dataset and unmirrored RH embedding is not visible anymore. Figure 5.3b shows the event selection efficiency when only applying the muon selection criteria except for the isolation requirement. From the good agreement and the small fluctuations, it can be deduced that the isolation component is the main contribution to deviations in the event selection efficiency.

Figure 5.2b shows the event selection efficiency when only applying the τ-lepton selection criteria. Here, all embedding methods show an over nPU mostly constant, between 10 % and 15 % higher selection efficiency than the validation dataset.

Figure 5.4a shows the event selection efficiency of the different embedding algorithms when all τ-lepton selection criteria but the requirement on the absolute isolation I^{τ} were applied. The almost perfect agreement between the embedding methods indicates that all significant discrepancies in the event selection efficiency are caused by the $\Delta\beta$-corrected isolation used for the τ-leptons. The distribution of the absolute amount of isolation from the tau discriminator used for the isolation

requirement, I^{τ}, is shown in Figure 5.4b. All embedding methods show a very similar trend towards better isolated τ-leptons than the validation dataset. This points to the MC simulation of the τ-lepton decay as a possible cause for the discrepancies.

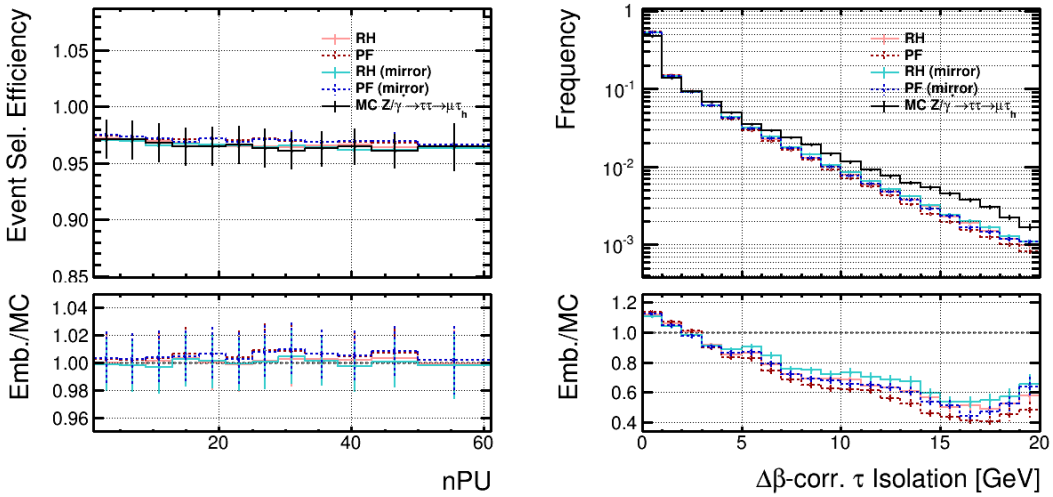


Figure 5.4: Figure 5.4a shows the event selection efficiency when applying all τ-lepton selection criteria but the isolation requirement. The almost perfect agreement between all five datasets indicates that the isolation component is the only source of significant deviations in the selection efficiency of τ-leptons. Figure 5.4b shows frequency of the absolute amounts of isolation used for the isolation requirement of τ-leptons. An increased number of events with isolation amounts below 2 GeV is visible and a lack of event with higher isolation contributions. The deviations are independent of the embedding method and the mirroring and cause the higher selection efficiency of the embedded datasets when applying the τ-lepton isolation criterion.

5.3 Kinematic properties

Muons

The distribution of kinematic properties of the muon in the $\mu\tau_{\mathrm{h}}$ final state after applying the baseline event selection is shown in Figures 5.5 to 5.7. The left plots show the kinematic properties of the unmirrored embedding methods compared to the validation dataset. The plots on the right side show the corresponding properties for the mirrored embedding methods.

Figure 5.5 shows the distribution of the azimuthal angle, ϕ, of the selected muons. In all embedding methods, the deviations are within the expected range of statistical fluctuations.

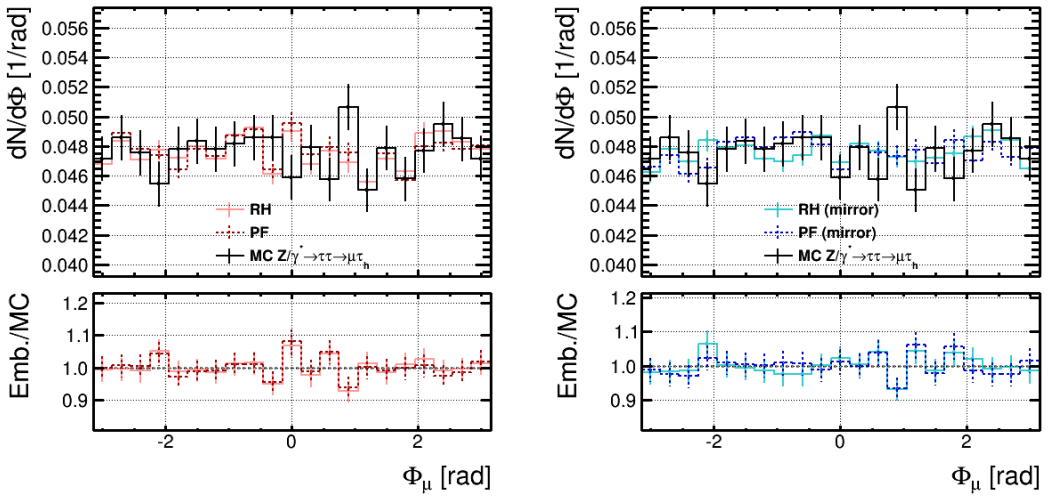


Figure 5.5: Distribution of the azimuthal angle, Φ, of the selected muons in the $\mu\tau_h$-channel. All embedding methods show a good agreement with the validation dataset.

The distribution of the pseudorapidity, η, of the selected muons is given in Figure 5.6. Due to the service chimneys of the cooling system of the solenoid, a lower number of muons around $|\eta| = 0.025$ is selected in the di-muon selection of the original events. Especially the mirrored embedding methods in the muon embedding showed a decreased number of events for these values of η. A compared to other areas of η reduced number of events around $|\eta| = 0.025$ is still visible but compared to the muon embedding, the effect is smeared out. This is caused by the emission of the neutrino in the τ-lepton decay that causes a deflection of the muons. The deviations in the η distribution are mostly below 5 %. A significant, systematic deviation in the embedding methods is not visible.

Figure 5.7 shows the distribution of the transverse momentum, p_T of the selected muons. All embedding methods show an excess of muons below 30 GeV. With increasing p_T fewer muons are reconstructed in the embedded dataset, especially above 60 GeV. Compared to τ-leptons in $Z \to \tau\tau$ decays, muons from $Z \to \mu\mu$ decays radiate a larger amount of final state radiation. This leads to on average lower amounts of transverse momentum for the embedded particles compared to a $Z \to \tau\tau$ dataset. Therefore, the bias towards lower transverse momentum is at least partly caused by the increased rate of final state radiation in the original events of the embedding. The up to 40 % reduced number of high energetic muons indicates that other effects e.g. a remnant of the di-muon selection could also contribute to this bias.

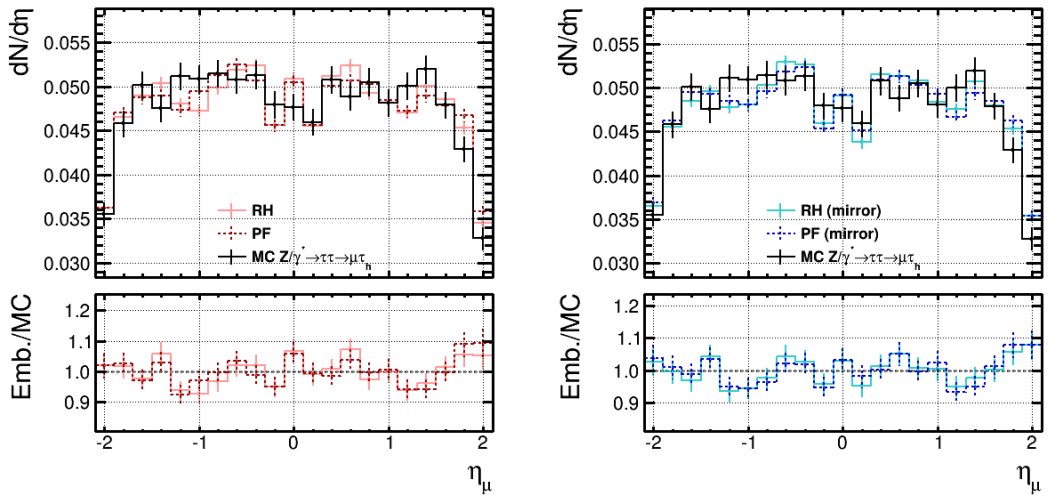


Figure 5.6: Distribution of the pseudorapidity, η of the selected muons in the $\mu\tau_{\mathrm{h}}$-channel. All embedding methods show a good agreement with the validation dataset.

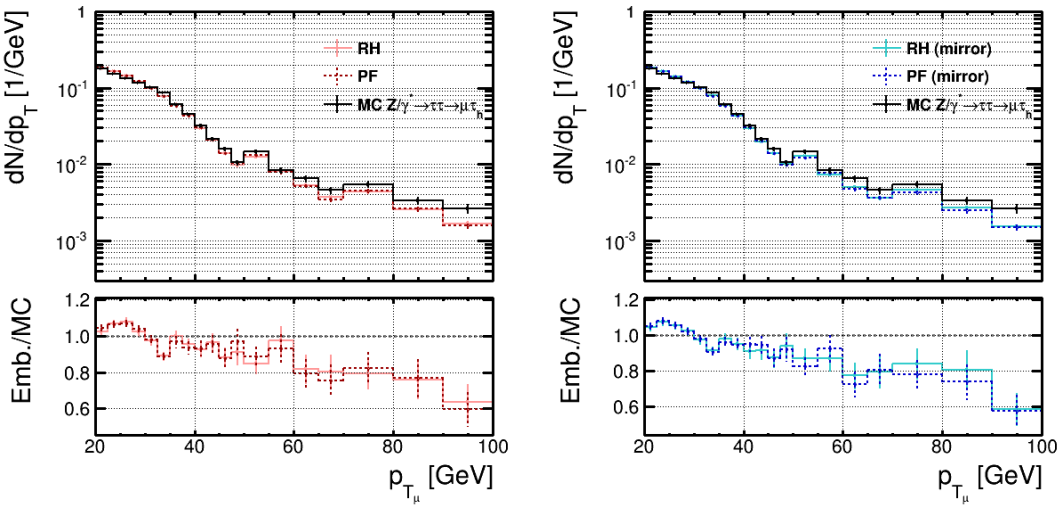


Figure 5.7: Distribution of the transverse momentum, p_{T}, of the selected muons in the $\mu\tau_{\mathrm{h}}$-channel. All embedding methods show an excess of muons with p_{T} below 30 GeV and a lack of high energetic muons, especially above 60 GeV. The shift towards lower p_{T} could be caused by the increased final state radiation in the $\mathrm{Z} \to \mu\mu$ events for the embedding.

Hadronically decaying τ-leptons

The distribution of kinematic properties of the hadronically decaying τ-leptons in the embedding and the validation dataset after applying the baseline event selection are shown in Figures 5.8 to 5.10.

As for the muons, the distribution of the azimuthal angle ϕ of the selected τ_h does not show systematic biases.

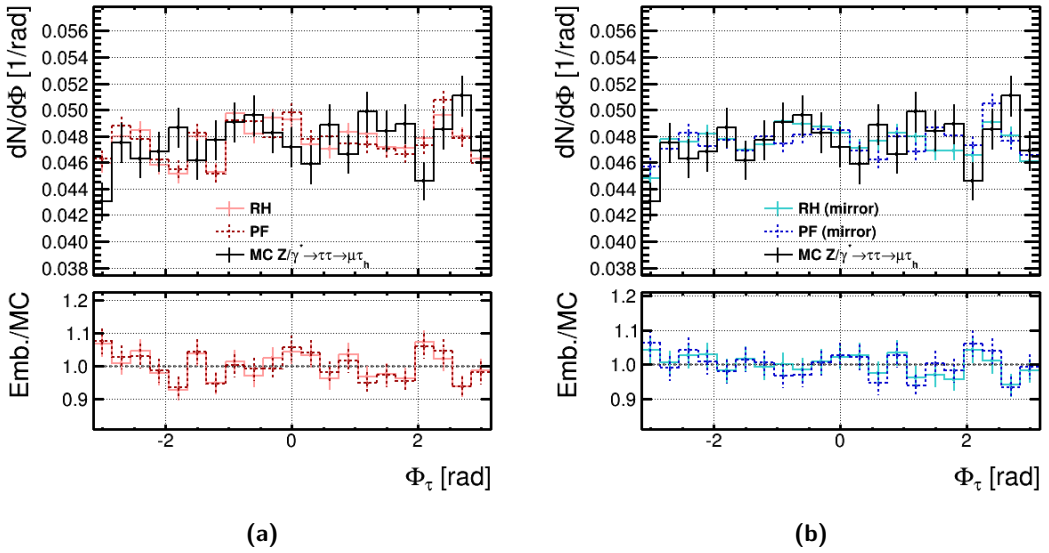


Figure 5.8: Distribution of the azimuthal angle, Φ, of the selected τ_h in the $\mu\tau_h$-channel. All embedding methods show a good agreement with the validation dataset.

The distribution of the pseudorapidity for the embedded dataset and the validation dataset are shown in Figure 5.9. The embedding methods show a good agreement within $|\eta| < 2$. Above this, the embedded datasets show an up to 15 % increased number of selected hadronically decaying τ-leptons.

The for the muons visible trend towards lower transverse momentum is less distinct for the reconstructed τ-leptons. Due to the hadronisation and the reconstruction from jets, the p_T resolution of reconstructed τ_h is lower than for reconstructed muons. Therefore, the effects that cause the trend to lower transverse momentum are smeared out and less significant in the reconstruction of hadronic decaying τ-leptons.

Di-τ mass

The neutrinos in the τ-lepton decay can carry away a large amount of transverse momentum that cannot be detected. This leads to a smearing our of the Z boson mass peak and a reduced combined visible mass, m_{vis}, of the $\mu\tau_h$-system. The reconstructed visible mass from the τ-lepton decays in the $\mu\tau_h$-channel, $m_{vis}^{\mu\tau_h}$, is shown in Figure 5.11. An increased selection efficiency of reconstructed di-lepton

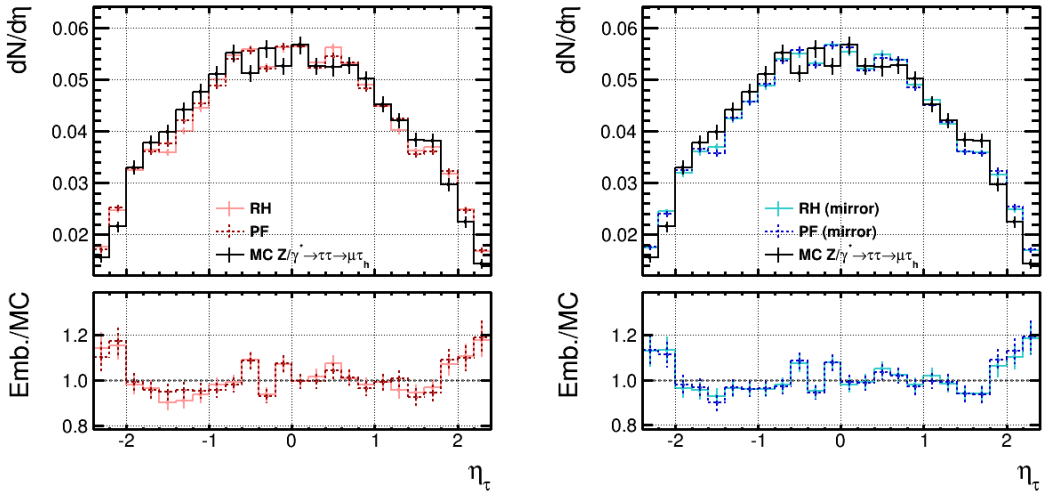


Figure 5.9: Distribution of the pseudorapidity, η of the selected τ-leptons in the $\mu\tau_{\mathrm{h}}$-channel. All embedding methods show a good agreement with the validation dataset within $|\eta| < 2$. The embedded datasets show a by approximately 15 % increased number of τ-leptons for $|\eta| > 2$.

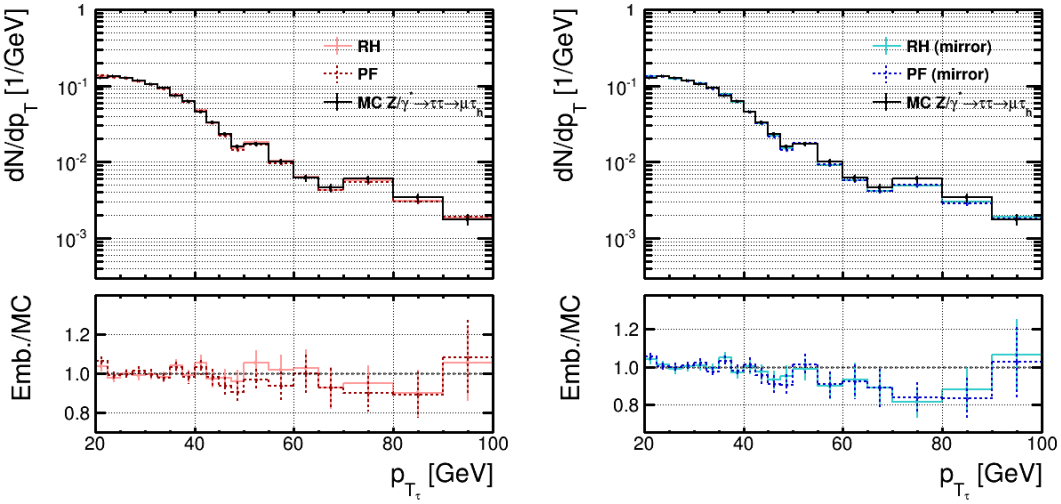


Figure 5.10: Distribution of the transverse momentum, p_{T}, of the selected τ_{h} in the $\mu\tau_{\mathrm{h}}$-channel. The shift towards lower p_{T} is less distinctive than for the reconstructed muons. Therefore, the overall agreement between the embedding datasets and the validation dataset is better for the reconstructed τ-leptons.

masses, as seen in the mirrored muon embedding around 60 GeV, is not visible in the $\mu\tau_h$-channel of the mirrored tau embedding. Below the Z boson mass, the visible mass of the reconstructed di-lepton system in all embedding methods shows a good agreement with the validation dataset. Above approximately 100 GeV, the number of reconstructed high energetic events is in all embedding methods by up to 30 % reduced.

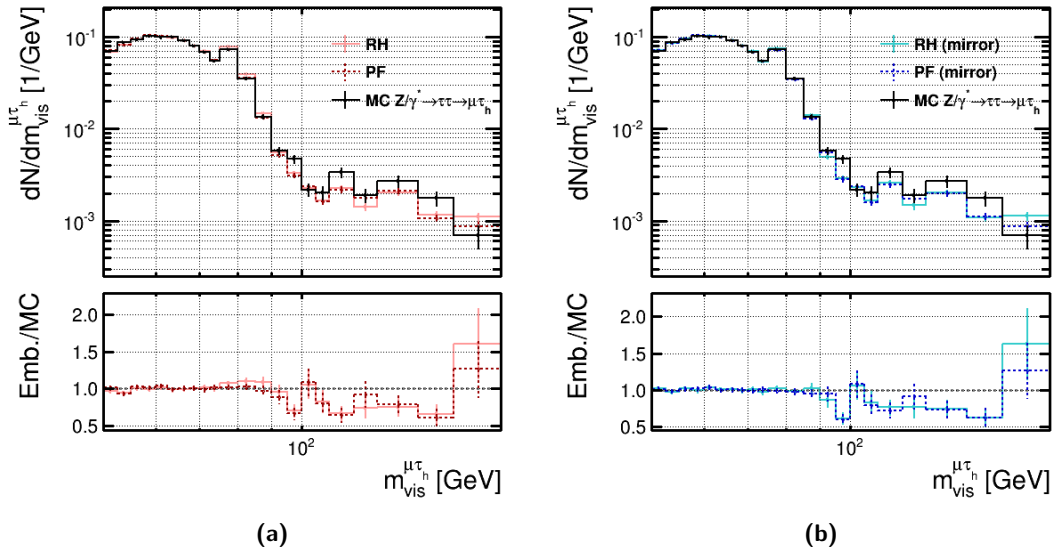


Figure 5.11: Distribution of the visible di-τ mass in the $\mu\tau_h$-channel for the unmirrored and mirrored embedding methods. Below the Z boson mass, the agreement of the validation dataset and the embedding methods is good. Above approximately 100 GeV, fewer events are selected in the embedding methods.

6 Conclusions and Outlook

The PF and the RH embedding algorithms were studied using the embedding of muons and τ-leptons. The embedding of τ-leptons was investigated exclusively in the $\mu\tau_{\mathrm{h}}$ final state of the di-τ decay. Both embedding algorithms were studied with and without the mirroring of the embedded particles as introduced in Section 3.4. All research was performed in version 7.0.7 of the software framework CMSSW.

Deviations observed in the event selection efficiencies are found to mainly originate from the application of isolation requirements in the embedded datasets. This was the case in both the muon embedding as well as the tau embedding.

Due to the requirement on the charged hadron isolation in the di-muon selection of $\mathrm{Z} \to \mu\mu$ events for the embedding, the unmirrored muon embedding methods show a 10 % higher selection efficiency than the $\mathrm{Z} \to \mu\mu$ validation dataset. When using the mirroring, this bias gets reduced to 2 %.

In the neutral hadron isolation component, a systematic bias is introduced in the mirrored PF embedding and the unmirrored RH embedding in presence of a muon. This bias is caused by the particle flow algorithm and its modifications to the neutral hadron particle collection if a muon was reconstructed. The deviations in the unmirrored RH embedding due to the incorrect handling of the modification increase with pileup up to 6 % for approximately 50 reconstructed pileup vertices. In the $\mu\tau_{\mathrm{h}}$-channel of the tau embedding, where only one muon is contained in the event, this effect is approximately half as large.

Deviations in the photon component were discovered to be mostly related to an improper handling of the final state radiation. Especially when mirroring, final state radiation contributions are lost, since the corresponding photons cannot be mirrored and their contribution to the isolation component is lost. Due to a partial correlation of muon final state radiation and the amount of isolation from photons close to the embedded particles, the effect of final state radiation can be suppressed by a requirement on a maximum amount of isolation within a small cone ΔR around the embedded particles.

In the tau embedding, the four different embedding methods show a 10 % to 20 % increase in selection efficiency compared to the validation sample. A dependence on the number of pileup vertices, nPU, cannot be excluded, but is expected to be smaller than 5 % for 50 pileup vertices in the $\mu\tau_{\mathrm{h}}$-channel of the tau embedding. The deviations due to the reconstruction and selection of the muons in the event are of the order of 7 % or less. The largest discrepancy in the tau embedding originates from the isolation of the τ-leptons. In all four embedding methods, the isolation of the τ-leptons is biased to lower values. This causes an approximately 10 % to 15 % increased selection efficiency of the baseline event selection in the embedded datasets.

Outlook

While a better understanding of deviations and their causes in the embedding was achieved, open questions worth studying still remain.

The largest deviations in the selection efficiencies of the studied embedding methods were introduced when applying isolation requirements. The charged hadron component of the isolation remains biased, even in the mirrored embedding samples. The remaining systematic deviation could be related to a lack of noise simulation in the embedded events or it could be a remnant from the di-muon selection. An embedded mirrored particle can e.g. still be close to the direction of one of the original muons in the event. Thus, some of the mirrored embedded particles still point into a direction that was considered in the di-muon selection of $Z \to \mu\mu$ events. This direction is expected to be biased to lower than average amounts of charged hadron isolation. These events would therefore cause the remaining increased selection efficiency. This idea could be tested by removing events where the isolation cones of the embedded and original particles overlap and comparing the charged hadron density profile of the remaining events.

The deviations due to the muon reconstruction in the neutral hadron component will probably be resolved with CMSSW 7.6. The part of the particle flow reconstruction algorithm that caused the discrepancies in the neutral hadrons in presence of a muon is expected to be removed from the algorithm from this software version on.

The origin of the discrepancies in the τ_h isolation needs to be studied further. Since all embedding methods independently of the mirroring show a very similar distribution of the isolation components, the origin is likely due to an aspect of the embedding common to all methods, for example the simulation of the di-τ decay in the simulated event of the embedding process.

Figure 6.1 shows the event selection efficiency in the $\mu\tau_h$-channel when removing only the isolation requirements for the muon and the hadronically decaying τ-lepton from the baseline event selection. This shows that when deviations due to the isolation components could be fully eliminated, only a deviation smaller than 2 % would remain. Therefore, the study of deviations in the isolation components should be one of the highest priorities for future embedding studies.

As shown before, the bias in the isolation is partly caused by requirements in the di-muon selection. Another possible bias from the di-muon selection is caused by its HLT requirement. Among others, this requirement is known to affect the pileup distribution and to be more efficient in certain regions of pseudorapidity. The HLT requirement is currently not recalculated in the embedding process. Due to the changed event topology, it cannot be used in the embedding. This raises the question, how large the impact of the HLT requirement is on the analysis and what the consequences were if it was removed from the di-muon selection.

The different amounts of emitted final state radiation from muons in the $Z \to \mu\mu$ decay is known to introduce biases in the simulated $Z \to \tau\tau$ decay of the embedding procedure. This will affect for example the reconstructed di-muon mass and the amount of photon isolation close to the reconstructed leptons. Additionally, the

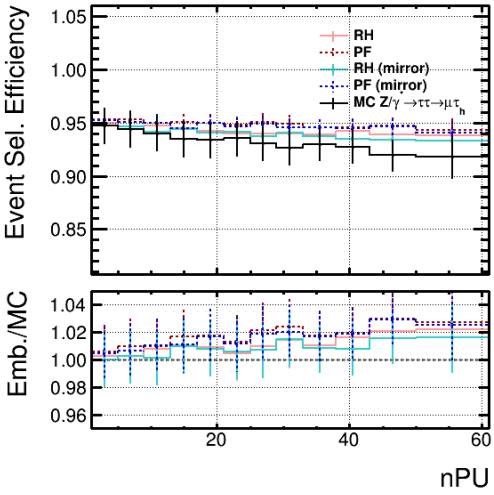


Figure 6.1: Event selection efficiency of the embedding methods without considering the isolation components. The event selection efficiency is significantly higher than with the isolation requirements and the deviations between the embedding algorithms and the validation dataset are smaller than 2 %.

mirroring transformation is sensitive to final state radiation as explained above. Since the exact effects and deviations due to the final state radiation are currently not known, a dedicated study of MC events without final state radiation could be performed. For this, a generator level muon final state radiation filter could be added to the di-muon selection and the validation dataset. This would allow for a study of the embedding without final state radiation contributions and would help to identify and quantify the impact of final state radiation in the various embedding methods.

If a dedicated MC event generation of $Z/\gamma^* \to \tau\tau$ and $Z/\gamma^* \to \mu\mu$ events for the embedding is performed, requirements on the minimum transverse momentum of the τ-leptons and muons e.g. of 20 GeV per lepton should be included. Additionally, the $Z/\gamma^* \to \tau\tau$ decays should be generated exclusively in the di-τ final states that are investigated. Thereby, the yield of the generated datasets could be improved. This is especially recommended due to the data storage needs of the RECO event format, needed for the embedding.

Additionally to the $\mu\tau_h$-channel, also the $e\tau_h$ channel of the di-τ decay should be studied to investigate effects and possible biases on the reconstruction of electrons.

A Input data generation with MC simulation

The datasets for the embedding as well as the validation dataset in this thesis were generated from an $Z/\gamma^* \to ll$ dataset[1]. The Drell-Yan process in this dataset was generated with a lower boundary on the invariant mass of the di-lepton system of 50 GeV. The simulation was done at a centre-of-mass energy of 13 TeV using the MC event generators madgraph and pythia8. The decay of the τ-leptons was simulated with Tauola.

To generate the input data for the embedding, this dataset was skimmed for $Z/\gamma^* \to \mu\mu$ events, using the MC generator level information of the events. This reduced $Z/\gamma^* \to \mu\mu$ dataset was then mixed with pileup.

The distribution of pileup is configured with two quantities, represented as vectors in CMSSW. One vector of the pileup mixing module called *probFunctionVariable* defines the expectation values, N, of a Poisson distribution $P_N(k)$ of superimposed pileup interactions. The vector *probValue* defines the corresponding probabilities, p, for each expectation value.

In the mixing of pileup, the expectation values N are chosen with the corresponding probability, p, and a random number of a Poisson shaped probability distribution $P_N(k)$ is pulled and taken as the amount of pileup interactions in the event. Therefore, when creating a pileup scenario that is flat in the probabilities p, the actual number of pileup interactions will be smeared out to lower and higher values than given in the flat probability distribution.

For the initial generation of the configuration file, an existing pileup scenario of CMSSW was used that is flat in p for $20 \leq N \leq 50$ with a bunch crossing rate of 25 ns. This scenario was then modified so all expectation values N from 0 to 49 had the same probability $p = 0.02$. The resulting pileup distribution is given in Figure A.1. The pileup events were taken from a so called *MinBias* dataset [2]. In this dataset, all known physics processes are simulated.

As next step, the detector simulation and reconstruction algorithms of the pileup-mixed $Z/\gamma^* \to \mu\mu$ dataset was run. The reconstructed events were then stored in the *RECO* event format that is needed of the RH embedding. This dataset corresponds to the validation dataset used in the muon embedding. The events used for the original events of the embedding were then retrieved by applying the di-muon selection on this reconstructed $Z/\gamma^* \to \mu\mu$ dataset.

[1] /DYJetsToLL_M-50_13TeV-madgraph-pythia8-tauola_v2/Fall13-POSTLS162_V1-v2/GEN-SIM
[2] /MinBias_TuneA2MB_13TeV-pythia8/Fall13-POSTLS162_V1-v1/GEN-SIM

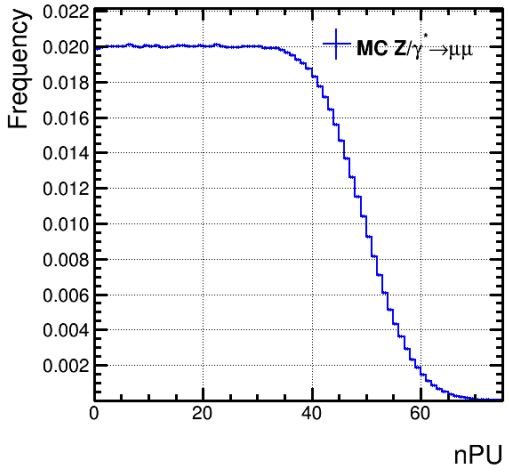


Figure A.1: Resulting pileup distribution for $p = 0.02$ from $0 \leq N \leq 49$. The flat pileup distribution is smeared to values above 49 pileup interactions.

The $Z/\gamma^* \to \tau\tau$ validation dataset for the tau embedding originates from the same $Z/\gamma^* \to ll$ dataset. This dataset was skimmed for $Z/\gamma^* \to \tau\tau$ events, mixed with pileup and reconstructed in the same way as the $Z/\gamma^* \to \mu\mu$ dataset of the muon embedding validation.

Bibliography

[1] ATLAS Collaboration, "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC", *Phys. Lett. B* **716** (2012) 1–29, DOI: 10.1016/j.physletb.2012.08.020 [arXiv:1207.7214].

[2] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett. B* **716** (2012) 30–61, DOI: 10.1016/j.physletb.2012.08.021 [arXiv:1207.7235].

[3] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* **13** (1964) 321–323, DOI: 10.1103/PhysRevLett.13.321.

[4] Peter W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Lett.* **12** (1964) 132–133, DOI: 10.1016/0031-9163(64)91136-9.

[5] Peter W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* **13** (1964) 508–509, DOI: 10.1103/PhysRevLett.13.508.

[6] G.S. Guralnik, C.R. Hagen and T.W.B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.* **13** (1964) 585–587, DOI: 10.1103/PhysRevLett.13.585.

[7] T. W. B. Kibble, "Symmetry Breaking in Non-Abelian Gauge Theories", *Phys. Rev.* **155** (1967) 1554–1561, DOI: 10.1103/PhysRev.155.1554.

[8] Peter W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons", *Phys. Rev.* **145** (1966) 1156–1163, DOI: 10.1103/PhysRev.145.1156.

[9] Robin G. Stuart, "On the Precise Determination of the Fermi Coupling Constant from the Muon Lifetime", *Nucl.Phys. B564 (2000) 343-390* (1999), DOI: 10.1016/S0550-3213(99)00572-6.

[10] Fabienne Marcastel, "CERN's Accelerator Complex. La chaîne des accélérateurs du CERN" (2013), URL: http://cds.cern.ch/record/1621583.

[11] Christiane Lefevre, *CERN LHC: The Guide*, Geneva, 2009, p. 60, CERN-Brochure-2009-003-Eng.

[12] *CMS detector design*, URL: http://cms.web.cern.ch/news/cms-detector-design (visited on 15/05/2015).

[13] CMS Collaboration, "CMS Tracking Performance Results from early LHC Operation", *Eur.Phys.J.C70:1165-1192,2010* (2010), DOI: 10.1140/epjc/s10052-010-1491-3.

[14] P. Adzicet, "Energy resolution of the barrel of the CMS Electromagnetic Calorimeter", *JINST* **2** (2007) P04004, DOI: 10.1088/1748-0221/2/04/P04004.

[15] CMS Collaboration, "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at sqrt(s) = 7 TeV" (2013), [arXiv:1306.2016].

[16] Victor Daniel Elvira, *Measurement of the Pion Energy Response and Resolution in the CMS HCAL Test Beam 2002 Experiment*, tech. rep., Geneva: CERN, 2004, URL: http://cds.cern.ch/record/800406.

[17] Shuichi Kunori, *CMS Hadron Calorimeter*, URL: https://indico.cern.ch/event/46651/contribution/8/material/slides/0.pdf (visited on 16/05/2015).

[18] CMS Collaboration, "Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7$ TeV", *JINST* **7** (2012) P10002, DOI: 10.1088/1748-0221/7/10/P10002, [arXiv:1206.4071].

[19] Florian Beaudette for the CMS Collaboration, "The CMS Particle Flow Algorithm", *Proceedings of the CHEF2013 Conference* (2014), [arXiv:1401.8155].

[20] CMS Collaboration, *Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons*, 2014, DOI: 10.1007/JHEP05(2014)104 [arXiv:1401.5041].

[21] ATLAS, *Status of Standard Model Higgs searches in ATLAS*, 2012, URL: https://indico.cern.ch/event/197461/contribution/1/material/slides/ (visited on 17/05/2015).

[22] CMS, *Status of the CMS SM Higgs search*, 2012, URL: https://indico.cern.ch/event/197461/contribution/0/material/slides/ (visited on 17/05/2015).

[23] CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", *Eur. Phys. J. C 75 (2015) 212* (2014), DOI: 10.1140/epjc/s10052-015-3351-7.

[24] Armin Burgmeier, "Position Resolution and Upgrade of the CMS Pixel Detector and Search for the Higgs Boson in the $\tau\tau$ Final State", PhD thesis, KIT Karlsruhe Institute of Technology, 2014, IEKP-KA/2014-06.

[25] CMS Collaboration, "Measurement of the properties of a Higgs boson in the four-lepton final state", *Phys. Rev. D 89 (2014) 092007* (2013), DOI: 10.1103/PhysRevD.89.092007.

[26] Johan Alwall et al., "MadGraph 5: Going Beyond", *JHEP* **06** (2011) 128, DOI: 10.1007/JHEP06(2011)128 [arXiv:1106.0522v1].

[27] Torbjörn Sjöstrand, Stephen Mrenna and Peter Skands, *PYTHIA 6.4 – Physics and Manual*, 2006, DOI: 10.1088/1126-6708/2006/05/026 [arXiv:hep-ph/0603175].

Hiermit versichere ich, die vorliegende Arbeit selbstständig verfasst und nur die angegebenen Hilfsmittel verwendet zu haben.

Benjamin Treiber
Karlsruhe, den 31. Juli 2015