Tyson, Richard (2023) *J/ψ near-threshold photoproduction off the proton and neutron with CLAS12.* PhD thesis.

https://theses.gla.ac.uk/83777/

J/ψ Near-Threshold Photoproduction off the Proton and Neutron with CLAS12

Richard Tyson

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Physics and Astronomy
College of Science and Engineering
University of Glasgow

May 2023

Abstract

In recent years, J/ψ photoproduction in the near-threshold region has seen a renewed theoretical interest due to the wealth of information it has to offer. Near-threshold J/ψ photoproduction proceeds through the exchange of gluons in the t-channel and is expected to provide unique insight about the nucleon gluonic gravitational form factors (GFFs). Previous studies at the Thomas Jefferson National Accelerator Facility (JLab) in Virginia, USA, have already measured the total and differential cross section of J/ψ near-threshold photoproduction on the free proton. This thesis presents the first measurements of J/ψ near-threshold photoproduction on the bound proton and bound neutron in a deuteron target.

The CEBAF Large Acceptance Spectrometer at 12 GeV (CLAS12) based at JLab's Hall B uses a 11 GeV electron beam impinging on a fixed liquid deuteron target. J/ψ is then produced via the exchange of a quasi-real photon and decays to a lepton pair which is detected alongside the recoil nucleon by CLAS12. Analysis procedures and in particular machine learning based techniques were developed in order to correctly identify final state particles and select J/ψ photoproduction events. The total and differential cross section of J/ψ near-threshold photoproduction on the bound proton and bound neutron were then measured from the detected J/ψ photoproduction events. The measurements of the cross sections on the bound proton and bound neutron agree well within their statistical uncertainty. This is consistent with the assumed two-gluon exchange production mechanism which is isospin invariant.

Overall a first measurement of the J/ψ near-threshold photoproduction cross sections on the bound proton and bound neutron was achieved. A better understanding of the mechanical properties of the nucleon, such as pressure and mass distributions, can be obtained by relating J/ψ production to the nucleon GFFs. This opens the way for exciting new insights into the internal structure of the nucleon, and in particular of the nucleon's gluonic content. An upcoming overhaul of the CLAS12 reconstruction will increase the reconstruction efficiency and the statistical precision of the preliminary measurements described in this thesis, with an expected gain in statistics of at least 50%. Future upgrades at JLab will allow to test some of the theoretical assumptions made in relating J/ψ near-threshold photoproduction to the nucleon GFFs.

Contents

List of Tables

List of Figures

Acknowledgements

To say that a lot of time and effort has gone into writing this thesis would be stating the obvious. Perhaps less obvious however is the pleasure that I took in working towards my Ph.D. thanks in no small part to the help, guidance and support of a number of individuals and organisations who I would like to thank.

First I would like to thank the U.K. Science and Technology Facilities Council for funding my research and allowing me to work towards my Ph.D.

I would also like to thank my colleagues in the Nuclear and Hadron Physics group. Thanks to my fellow Ph.D. students for always answering my trivial questions and making any conference or JLab trip a fun experience. Thank you to the regular attendees of the Friday morning JLab meetings who always had insightful advice. In particular, I'm very grateful to Derek Glazier for all his help whether with using his extremely helpful software or for his ever pertinent suggestions. Many thanks to David Ireland for his supervision, continued support and insight who helped make my Ph.D. a productive and enjoyable endeavour. Finally, I cannot overstate my gratitude for Bryan McKinnon whose invaluable guidance was key in helping me make the most of this opportunity and produce the results that are presented in this thesis.

This work would not have been possible without the CLAS12 Collaboration. It took the collaborative effort of many dedicated scientists to produce and calibrate the data that I used for my thesis, and for that I'm very grateful. In particular I would like to thank Gagik Gavalian for his guidance with AI based projects and Silvia Niccolai for her continued support and key role in coordinating the RG-B experiment. I am also extremely grateful to Stepan Stepanyan, Valery Kubarovsky, Rafayel Paremuzyan, Joseph Newton and Pierre Chatagnon for the many helpful discussions on the J/ψ analyses at CLAS12.

Finally, to my friends and family close and far, a huge thank you. A ma maman et à mes frères, merci pour toutes vos pensées et gentils mots qui m'ont gardé motivé. Thanks Dad for all the support from Berlin! To my furry friend Annabelle, thank you for alerting me to any suspicious noise that could derail my work and for always reminding me to take a break. To my dear Lauren, thank you for all your kindness, support and patience whenever I told you I just had five more minutes of work to do. You made these past few years full of joy and I can't ever thank you enough for that!

And to the reader and those who gave me feedback, thank you for your reading my thesis!

Declaration

With the exception of chapters 2, 3 and 4 which contain introductory material, all the work in this thesis was carried out by the author unless otherwise explicitly stated.

Chapter 1

Introduction

Nuclear physics is the field of physics that studies the nuclei of atoms and their constituents, namely protons and neutrons. Particle physics is the field of physics that studies the particles and forces that constitute matter and radiation. At the interplay between both fields, hadronic physics studies hadrons, composite particles made of elementary particles named quarks bound together by gluons that mediate the strong force. In the quark model, hadrons are made of two quarks in the case of mesons, three quarks for baryons, or four or five quarks for tetraquarks or pentaquarks. Protons and neutrons are baryons, hadrons with three quarks. Hadronic physics therefore concerns itself with the study of Quantum Chromodynamics (QCD), the theory describing the strong interaction that binds quarks and gluons into hadrons.

The standard model (SM) of particle physics describes three of the four known fundamental forces of the universe, namely the electromagnetic force, the weak force and the strong force. Electromagnetism is an interaction that takes place between particles with electric charge, mediated by an electromagnetic field. Every force within the SM is mediated by bosons, particles with integer spin. The electromagnetic force is mediated by photons, massless bosons better known as the physical quanta of light. Quarks also interact electromagnetically due to their electric charge, and the quark content of the nucleon can therefore be probed using photon beams. The weak force describes the interaction between subatomic particles that is responsible for the radioactive decay of atoms such as nuclear fission and nuclear fusion. The weak force is mediated by the heavy massive $W^{\pm}$ and Z^0 bosons. Finally, as mentioned earlier, the strong force is mediated by gluons and binds quarks into hadrons. Quarks are fermions, meaning that they have half-integer spin, and come in three generations. The lightest quarks are called up (u) and down (d). The up-quark carries an electric charge of $+\frac{2}{3}$ and the down-quark $-\frac{1}{3}$. The two heavier versions of the up-quarks are called charm (c) and top (t) and also have charge $+\frac{2}{3}$, while the two heavier versions of the down-quarks are called strange (s) and bottom (b) with charge $-\frac{1}{3}$. Anti-quarks, denoted by a bar such as $\bar{c}$ for anti-charm, have opposite properties to their quark equivalents. The strong force counterpart of the electric charge of the quark content of the nucleon is called colour charge, which is where QCD takes its name. Similarly to how quarks can have positive of negative charge, quarks can have one of three colours, blue, red or green, and anti-colours, anti-blue, anti-red or anti-green. Gluons, although they do not have an electric charge, carry a colour charge, meaning that they can self-interact.

The colour charge of quarks and gluons leads directly to two fundamental properties of QCD, namely confinement [1] and asymptotic freedom [2,3]. Confinement is a consequence of the fact that all hadrons are colour neutral. Mesons are therefore composed of a quark anti-quark pair, both with the same colour anti-colour that cancel each other out to stay colour neutral. Baryons will have a quark of each available colour, which leads to their colour neutrality. This simple fact leads to a profound aspect of QCD; quarks are never observed on their own and are always confined within a hadron. This is related to the fact that the strong coupling constant α_s (which is not constant) that describes the magnitude of the coupling between quarks and gluons increases as the distance between quarks increases. Pulling two quarks apart would increase the binding energy between the two quarks to the point where it would be more kinematically favourable to create a new quark anti-quark pair from the vacuum. As opposed, asymptotic freedom is a consequence of the fact that at small distances, α_s becomes small, leading to quarks appearing as almost free when close to one another. The magnitude of α_s has a direct consequence for theoretical predictions in QCD, as at small α_s perturbative theory can be used to make such predictions, whilst at large α_s perturbative theory cannot be used and other more complicated calculations are then used. This defines the perturbative or non-perturbative regions. It should also be pointed out that the energy of the probe used to study QCD can be thought of as a magnifying glass, with high energies probing small distances and vice-versa.

Different quark combinations affect the quantum numbers of the hadrons that they form. For example, the proton contains two up-quarks and a down-quark and therefore has an electric charge of one. Neutrons on the other hand are composed of one up-quark and two down-quarks and therefore has an electric charge of zero. Protons and neutrons are further called isospin partners based on the up and down quark content of the protons or neutrons, with the proton having isospin $\frac{1}{2}$ and neutrons having isospin $-\frac{1}{2}$. Isospin partners are affected equally by the strong force. More generally, mechanisms that do not depend on the isospin of a particle are called isospin invariant. The two, three, four, five and so on, quarks that determine the hadrons quantum numbers are called valence quarks. Any hadron may contain an additional indefinite number of virtual sea quarks and gluons which do not influence the hadron's quantum numbers. A virtual particle is a transient particle that exhibits some of the characteristics of an ordinary particle, while having its existence limited by the uncertainty principle. For example, based on the uncertainty principle for energy and time $\Delta E \Delta t \geq \frac{\hbar}{2}$, a virtual particle can borrow an amount of energy ΔE from the vacuum allowing it to exist for an amount of time Δt. Virtual particles are often used to describe the interaction between two real particles, where for example the interaction of an electron with a nucleon can be described via the exchange of a virtual photon. Virtual particles do not necessarily carry the same mass as the corresponding real particle, although they always conserve energy and momentum [4]. For example, a virtual photon does not necessarily have a mass of zero, and the closer to zero its mass, the less virtual or more real it is. It is also worth noting that a virtual particle is not any less real than an ordinary particle. They are only called virtual as they are temporary in the sense that they cannot be directly detected but they affect physical quantities in measurable ways.

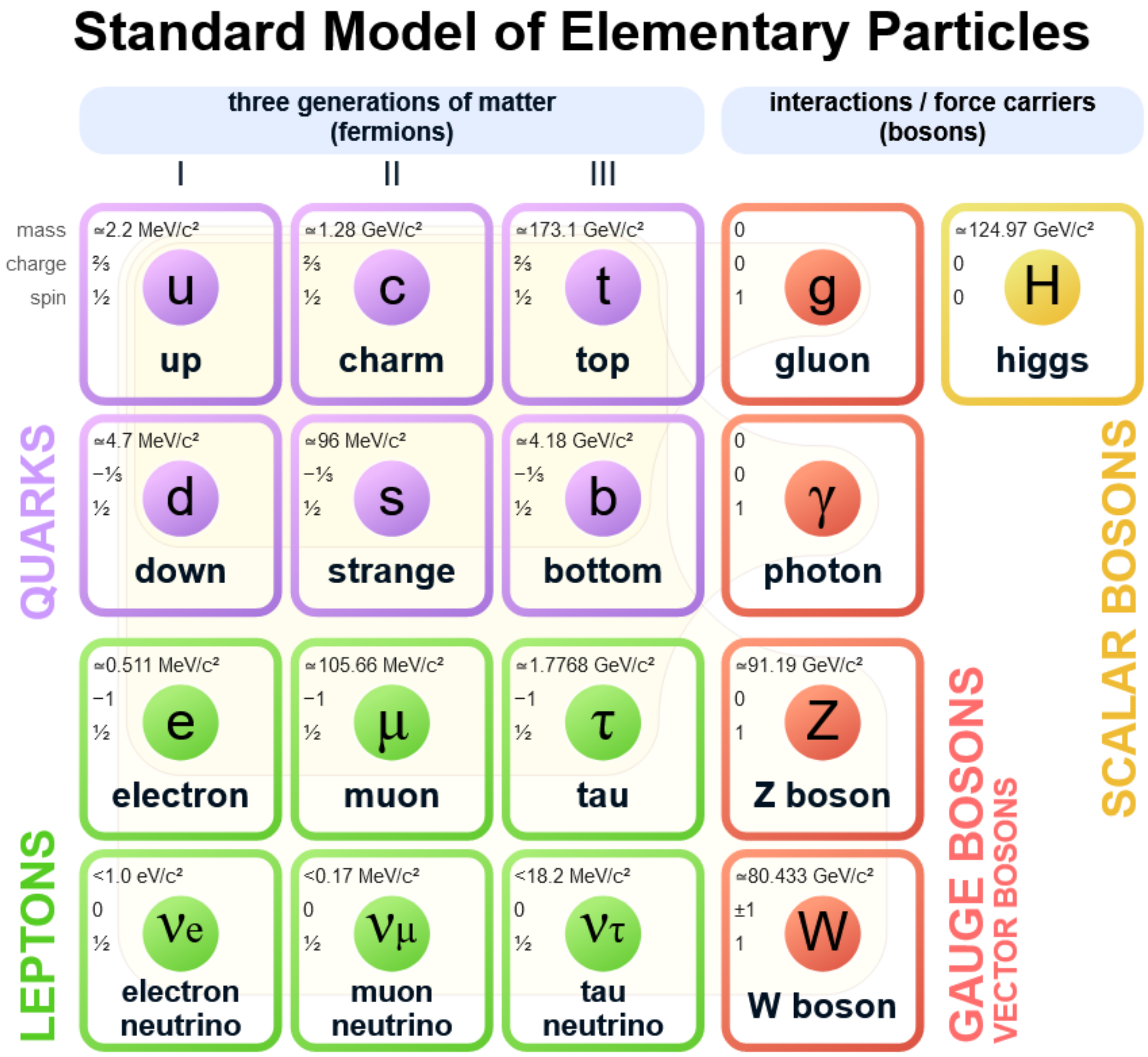


Figure 1.1: The Standard Model of Particle Physics. Taken from [5,6].

Figure 1.1 shows a recap of the standard model of particle physics. The last two particle types left to describe are the Higgs boson and leptons. The Higgs boson can be seen as a manifestation of the Higgs field which plays an important role in generating the mass of fundamental particles in the SM such as quarks and leptons. Leptons are fundamental particles that do not interact through the strong force. These are also fermions, meaning that they have half integer spin, and come in three generations as well. The lightest lepton is the electron which has a charge of -1, with its heavier partners the muon and tau also having an electric charge of -1. Their almost massless counterparts, neutrinos, are chargeless and can only interact through the weak force.

A glaring omission from the SM of particle physics is gravity, the fundamental interaction that causes the attraction between any two massive systems. The gravitational force finds its origin in a second-order tensor called the energy momentum tensor [7] with the assumed force carrier for gravity called the graviton, a spin-2 boson. The energy momentum tensor describes the density and flux of energy and momentum in spacetime [7]. This density and flux of energy and momentum are the sources of the gravitational field in the Einstein field equations of general relativity [7], just as mass density is the source of such a field in Newtonian gravity. One particularity of QCD as noted above is the fact that the strong coupling constant varies with scale. Another consequence of this scale dependence is that the trace of the energy momentum tensor is non-zero, unlike classical field theory which is invariant under scale transformations. This non-vanishing trace is called the trace anomaly of QCD and has a direct effect on the mass of the nucleon. The energy momentum tensor also contains the Hamiltonian of the nucleon, the operator that measures the energy of a state and so determines the rest mass of the nucleon. This is interesting in the context of the nucleon mass, which is sometimes referred to as puzzling. Indeed, as mentioned above, the Higgs field is responsible for generating the quark rest masses. However, if the mass of the two up-quarks and the one-down quark are summed, this does not add up to the mass of the proton of about 938 MeV. In fact, the nucleon derives its mass from the quark-gluon dynamics of its underlying structure [8]. The nucleon mass can then be decomposed into the quark energy contribution M_q, the gluon energy contribution M_g, the quark mass contribution M_m, and the trace anomaly contribution M_a [8]. The magnitude of the contribution of these terms to the proton mass are quantified by the QCD trace anomaly parameter b and the momentum fraction carried by all quarks a such that:

$$\begin{aligned} M_q &= \frac{3}{q}(a - \frac{b}{1+\gamma_m})M_N \\ M_g &= \frac{3}{4}(1-a)M_N \\ M_m &= \frac{4+\gamma_m}{4(1+\gamma_m)}bM_N \\ M_a &= \frac{1}{4}(1-b)M_N \end{aligned} \tag{1.1}$$

where M_N is the mass of the nucleon and γ_m is the quark mass anomalous dimension, which can be calculated perturbatively in QCD and describes the renormalization information used in the

QCD calculation [8]. This decomposition of the different components of the nucleon mass is referred to as the nucleon mass decomposition or sometimes Ji's decomposition [8]. At present the QCD trace anomaly parameter b is not well constrained, and as such it is hard to estimate the energy momentum tensor contribution to the nucleon mass or to calculate the nucleon mass from first principles. However, J/ψ photoproduction is expected to shed light on these important outstanding problems in QCD.

J/ψ, a meson composed of $c\bar{c}$ valence quarks, has played an important role in modern day particle and nuclear physics since its discovery in 1974 [9, 10] that started the November Revolution [11, 12] which followed the confirmation of a fourth quark flavour, charm. J/ψ was discovered independently at the Stanford Linear Accelerator Center [10], with this measurement headed by Burton Richter, and at the Brookhaven National Laboratory [9] by Samuel Ting and his collaborators. J/ψ owes its composite name to this particularity, with both laboratories choosing a different name and the community adopting both out of fairness. What was striking about the discovery of J/ψ was that it cemented the existence of quarks, which was still debated at the time. The point like structure of the constituents of hadrons had been studied at SLAC but the emergence of a new quark flavor, quickly solidified by the discovery of the excited states of J/ψ and charmed mesons such as as D and $\overline{D}$, would eventually lead to the Standard Model of Particle Physics as it is now known. All of these developments led to the establishment of QCD.

One might wonder why J/ψ production remains relevant, given the wealth of measurements from collider experiments that followed the November Revolution. The answer is twofold: First, theoretical predictions on the production mechanism of J/ψ near its photoproduction threshold of 8.2 GeV reveal how J/ψ can be used to probe the gluonic contents of the nucleon, which itself is intimately linked to the proton and neutron mass decomposition [8] as discussed above. Secondly, J/ψ photoproduction, whereby J/ψ is produced by the interaction of a photon with a given target, has not been extensively studied in the past near the production threshold. However, recent upgrades at the Thomas Jefferson National Accelerator Facility (JLab) now allow for detailed studies of J/ψ photoproduction in this near-threshold region. As such, J/ψ photoproduction at JLab can seemingly contribute to a better understanding of the structure of nucleons, which in turn could help shed light on the nucleon mass puzzle.

The basis for this is rooted in the equivalence between a spin-2 system and the graviton. Any spin-2 field would give rise to a force indistinguishable from gravity [7] as a spin-2 field will couple to the energy momentum tensor. This allows to estimate the nucleon gravitational form factors (GFFs) that encode mechanical properties of the nucleon such as shear, mass or pressure distributions [13]. The quark GFFs have already been estimated in the context of Deeply Virtual Compton Scattering (DVCS), where in-going and out-going spin-1 photons conspire to create a spin-2 system which allows to probe the pressure [14] and shear [15] distributions of the quark content of the nucleon. The inverse process to DVCS, called Timelike Compton Scattering (TCS), may also help study the quark GFFs. Similarly, J/ψ photoproduction includes processes involving the exchange of two spin-1 gluons, mimicking a spin-2 system such as a graviton [16]. In such a way the interaction couples to mass similarly to a gravity interaction, allowing to probe the gluonic GFFs. Due to the extreme weakness of the gravitational interaction the direct measurement of the gravitation field produced by a single proton is clearly impossible [13]. Likewise, studying graviton-proton scattering is currently off limits for current experiments. As such, DVCS, TCS and J/ψ photoproduction offer an unrivaled approach to studying the mass properties of the nucleon whilst giving a new direction to hadronic physics. Given that protons and neutrons form up to 99% of visible matter, a detailed understanding of their mass properties would be invaluable to modern day nuclear and particle physics.

Chapter 2

J/ψ Near-threshold Photoproduction

To place the rest of the work presented in this thesis into context, this chapter will start by giving an overview of J/ψ photoproduction before describing the theoretical predictions for the observables that can be measured from J/ψ near-threshold photoproduction. Finally the specifics of the measurements described in this thesis will be discussed in the final section.

2.1 J/ψ Quasi-real Photoproduction

The aim of this section is to give a basic overview of J/ψ photoproduction. The experiment considered in this thesis used a fixed target containing either free or bound nucleons. An additional caveat is added here due to the electron beam used at the CEBAF Large Acceptance Spectrometer (CLAS12). Instead of considering J/ψ photoproduction, the quasi-real photoproduction scenario will be presented here. In this scenario, the interaction of an electron beam with the target nucleon is mediated by a quasi-real photon, meaning that the mass of the virtual photon is nearly zero. The electron beam can then be described by an equivalent beam of quasi-real photons based on the Equivalent Photon Approximation (EPA) [17, 18].

A diagram of J/ψ quasi-real photoproduction on a proton target is shown in Figure 2.1. The electron beam interacts with the proton via the exchange of a quasi-real photon and produces J/ψ. The virtual mass of the photon is equal to Q^2, the energy transfer to the scattered electron squared. Quasi-real photoproduction therefore occurs at low Q^2. The four-momentum transfer to the recoil proton is measured by the Mandelstam variable t.

After the interaction shown in Figure 2.1, J/ψ will decay to a lepton pair of opposite charge, l^+l^-, where the lepton pair can either be a di-electron or di-muon pair. The decay of J/ψ to hadrons for states lighter than the charged D meson is suppressed according to the OZI rule, independently formulated by Okubo, Zweig and Iizuka [19–21]. For such hadrons, the decay must proceed only via virtual gluons mediating the interaction, each of which must have enough

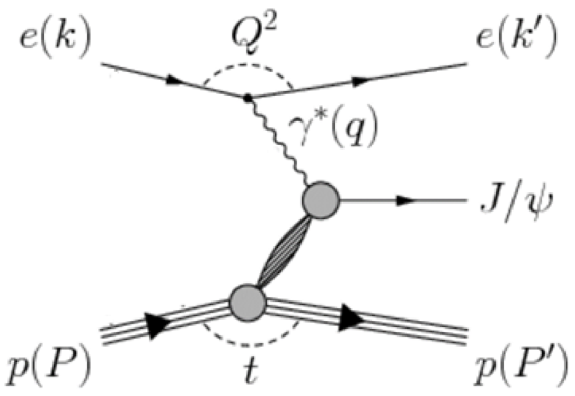


Figure 2.1: A diagram of J/ψ quasi-real photoproduction on the proton. A virtual photon γ^* with four-momentum q mediates the interaction between an electron e with four-momentum k and a proton p with four-momentum P. The proton is represented here by three lines denoting the three valence quarks of the proton. The interaction produces J/ψ. t represents the four-momentum transfer to the recoil proton p with momentum P'. Q^2 represents the squared energy transfer to the scattered electron e with momentum k' and is equal to the virtual mass of the photon γ^*. Adapted from [22].

energy to produce a quark-antiquark pair. As the coupling constant of QCD α_s decreases with energy or momentum transfer, it will therefore appear small to the gluons mediating the decay which in turn will be suppressed. The OZI rule was known by the time of the J/ψ discovery. This in turn helped to cement the existence of the charm quark as the initial measurements of J/ψ indicated that J/ψ had a longer lifetime than lower mass states. This longer lifetime is due to the OZI rule which suppresses certain decays, notably that to pions made of up and down quarks. Only a fourth quark flavour could explain the long lifetime J/ψ, as the ϕ meson made of bound $s\bar{s}$ quarks had already been discovered.

Quantities like Q^2 and t shown in Figure 2.1 can be calculated from the four-momenta of the final state particles and the known four-momenta of the beam and target. These calculations are made in natural units, with the speed of light $c = 1$ and the reduced Planck constant $\hbar = 1$. For example, for $\mathbf{k}'$ the four-momentum of the scattered electron and $\mathbf{k}$ that of the incoming electron beam, Q^2 can then be calculated as:

$$Q^2 = -(\mathbf{k} - \mathbf{k}')^2 \tag{2.1}$$

For $\mathbf{P}$ and $\mathbf{P}'$ the four-momentum of the target nucleon and recoil nucleon respectively, t can then be calculated as [23]:

$$t = (\mathbf{P}' - \mathbf{P})^2 \tag{2.2}$$

The energy of the virtual photon can be calculated from the energy of the final state leptons and nucleon, $E_{l^\pm}$ and $E_{N'}$ respectively, and the energy of the target nucleon E_N. Given that the target is at rest, and that the mass of the nucleon is written as M_N, the energy of the final virtual photon simplifies to:

$$E_\gamma = E_{l^+} + E_{l^-} + E_{N'} - M_N \tag{2.3}$$

The invariant mass of the J/ψ decay to a lepton pair (called IM here) can be calculated from the four-momentum of the lepton pair $\mathbf{l}^\pm$ [23]:

$$IM = \sqrt{(\mathbf{l}^+ + \mathbf{l}^-)^2} \tag{2.4}$$

The invariant mass of a di-lepton pair produced in the J/ψ decay will be the mass of J/ψ, 3.097 GeV. J/ψ production can then be identified by a peak at the J/ψ mass in the di-lepton invariant mass spectrum. The missing mass squared (MM^2) of the reaction is taken as the difference between the four-momentum of the beam and target, and the final state particles [23]:

$$MM^2 = (\mathbf{N} + \mathbf{k} - \Sigma_i^N \mathbf{p_i})^2 \tag{2.5}$$

for $\Sigma_i^N \mathbf{p_i}$ the sum of the four-momentum of all final state particles. The missing mass of the reaction should then be zero if all the final state particles are detected, as any reaction will conserve energy and momentum. If the missing mass squared was calculated using all but one of the final state particles then it would be equal to the square of the mass of the missing particle. For the analyses of J/ψ photoproduction at CLAS12 presented in this thesis the scattered electron goes undetected, as will be explained in the last section of this chapter. In this case the missing mass squared should peak at the mass squared of the electron, which is roughly zero at GeV energies. The missing mass squared can therefore be used to check that a given reaction is occurring as it will peak at a given mass unless the wrong assumption is made on the reaction that produced the particles used in the calculation.

All variables described above are then used for the measurement of J/ψ quasi-real photoproduction at CLAS12. The invariant mass spectrum is fitted to give the number of J/ψ produced in the experiment. Applying a threshold to the missing mass squared allows to select exclusive J/ψ events, and an upper limit on Q^2 allows to select quasi-real photoproduction events. Binning the cross section in E_γ allows to disentangle near-threshold J/ψ production from J/ψ production at higher photon energies. Furthermore, as will be detailed in the next sections, the production mechanism of J/ψ quasi-real photoproduction can be studied by calculating the total cross section as a function of E_γ, whilst the differential cross section as a function of $-t$ is directly connected to the nucleon mass radius and the gluon gravitational form factors.

2.2 Vector Meson Dominance and the J/ψ Production Mechanism

Before moving on to a discussion of the J/ψ production mechanism and how J/ψ can be used to probe the gluonic content of the nucleon, some time should be spent describing the basic theoretical framework underlying these predictions. First introduced by J.J. Sakurai in the 1960s [24], vector meson dominance (VMD) describes the interaction between photons and hadronic matter, which can also be applied to electron beams based on the EPA.

VMD originated from several observations: First, rather unexpectedly, the photon-hadron interactions bore notable similarities to hadron-hadron interactions [25], especially when considering the behavior of the total cross section. Both show strong resonances, in particular neutral vector mesons such as ρ, ω or ϕ, with the photon total cross sections smaller than the hadronic ones by approximately the fine structure constant α_{em} in order of magnitude [25]. The fine structure constant α_{em} is also known as the electromagnetic coupling constant. Similarly, the photon interaction does not seem to depend primarily on the charge of the target as the total cross sections on neutrons and protons are nearly the same [25]. Both of these observations point to the fact that photons do not solely interact with a hadron's electric charge.

The suggestion from Sakurai was to instead describe the physical photon $|\gamma\rangle$ as a superposition of a bare photon $|\gamma_B\rangle$ and a hadronic component $|h\rangle$ of order $\sqrt{\alpha_{em}}$ [25]. At high energies the bare photon accounts for a small portion of the interaction whereas the hadronic component undergoes conventional hadronic interactions. The physical photon involved in the interaction can therefore be expressed as:

$$|\gamma\rangle \cong \sqrt{Z_3}\,|\gamma_B\rangle + \sqrt{\alpha_{em}}\,|h\rangle \tag{2.6}$$

where Z_3 is introduced to assure the proper normalisation of $|\gamma\rangle$ [25]. The hadronic components of the photon, $|h\rangle$, must then have the same quantum numbers as the photon, namely $J^{PC} = 1^{--}$. This corresponds to vector mesons. The large photoproduction cross-section of the ρ, ω or ϕ mesons suggest that they provide important contributions to $|\gamma\rangle$. The more restrictive assertion that these three mesons are the sole hadronic constituents of the photon, and that the bare photon component cannot interact with hadrons is referred to as VMD. The less restrictive assumption that $|h\rangle$ has more constituents than ρ, ω or ϕ is referred to as generalized vector dominance (GVD).

As described above, in the conventional VMD framework, the photoproduction (γ) of a vector meson (V) on a target nucleon (N) can be related to the VN scattering as [26]:

$$\frac{d\sigma_{\gamma N\to VN}}{dt} = \kappa\frac{3\Gamma(V\to e^+e^-)}{\alpha_{em}m_V}\frac{d\sigma_{VN\to VN}}{dt} \tag{2.7}$$

where κ is a kinematic factor, Γ is the partial decay width of the vector meson to e^+e^- and α_{em} and m_V are the fine structure constant and the meson mass respectively. A diagram of this process is shown in Figure 2.2. The photon γ fluctuates into a quark anti-quark pair $q\bar{q}$ of transverse size $r_\perp$ at a distance l_C before the interaction named the coherence length. The $q\bar{q}$ pair then scatters off the nucleon with impact parameter b. The $q\bar{q}$ pair finally forms a vector meson V at some distance after the scattering called the formation length l_F [27].

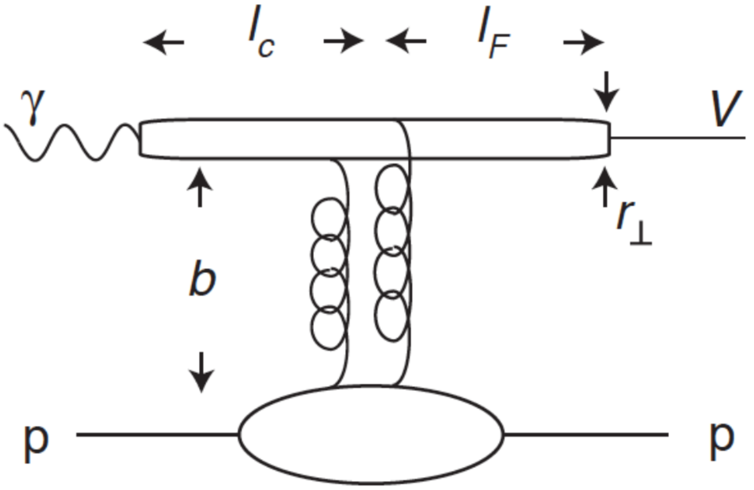


Figure 2.2: Diagram of vector meson photoproduction in the VMD picture, taken from [27].

For J/ψ photoproduction near-threshold, the coherence length will be small, of order $l_C \approx 0.36$ fm [27] with the formation length of order $l_F \approx 1-2$ fm [29]. The large mass of J/ψ imposes a small transverse size, $r_\perp \approx 0.13$ fm [27]. The minimum momentum transfer t_{min} is large such that $t_{min} \approx -1.7$ GeV^2 [27]. Therefore all of the target quarks have to transfer their energy to the charm quarks withing a creation time of $\frac{1}{m_c}$, for m_c the mass of a charm quark [27]. The target quarks must also be within $r_\perp$ of the $c\bar{c}$ pair and each other [27]. Charmonium ($c\bar{c}$) production near-threshold must therefore occur at small impact parameters $b \approx \frac{1}{m_c} \approx 0.13$ fm [27].

At high energies, the dominant contribution to J/ψ photoproduction comes from so called leading-twist diagrams, where only one quark from each of the colliding particle participates in the reaction [27]. Higher-twist, multi-quark processes are suppressed, due to a requirement for all involved quarks to be in area of $\frac{1}{m_c^2}$. Twist here refers to the difference of the dimensionality and spin of the production mechanism [30]. For charmonium photoproduction near-threshold, in order to have elastic scattering where the nucleon stays intact, quarks in the nucleon must share the large momentum transferred to the $c\bar{c}$ pair and must be in a compact Fock state. The effective proton radius and lifetimes of this compact Fock state then reduce to the lifetimes and scales mentioned in the previous paragraph. In that regime, the leading-twist processes cease to dominate the higher-twist processes [27]. For each quark involved in charmonium production a single gluon is exchanged, as shown in Figure 2.3 a). The VMD based theoretical models then predict that near-threshold, higher-twist two-gluon exchange or three-gluon exchange, shown in Figure 2.3 b), start to dominate charmonium production, with three-gluon exchange dominating the two-gluon exchange as the photon energy decreases to the J/ψ production threshold.

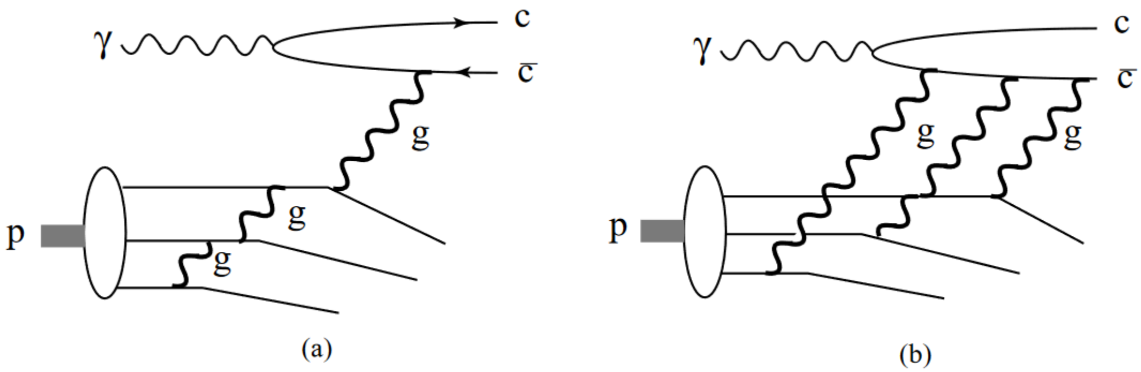


Figure 2.3: a) Diagram of leading-twist single gluon exchange. b) Diagram of three-gluon exchange. Both taken from [27].

It should be noted that VMD was originally introduced before the development of QCD to describe electromagnetic interactions of light hadrons. It remains popular to date as it allows for a non-perturbative reaction theory that can explain quark and anti-quark scattering from hadron targets into vector-meson final-states. However, QCD based approaches have started to emerge for some given problems. Similarly, holographic QCD [31] gives approximate analytical solutions to non-perturbative problems by mapping QCD to a higher-dimensional space. As such, although VMD tends to be the favoured model for hadronic physics, other theoretical frameworks can also be employed. A good review of the different models for the J/ψ production mechanism and their underlying theoretical framework is given in [32].

2.2.1 An aside on the LHCb Pentaquarks

In 2015 [33] and 2019 [34], the LHCb collaboration announced the discovery of candidates for pentaquark states with hidden charm, denoted P_c^+. These pentaquark candidates would then contain five quarks with a charm anti-charm pair corresponding to their hidden charm components. The pentaquark candidates were identified in the $J/\psi p$ invariant mass spectrum, shown in Figure 2.4. As such, P_c^+ decay could provide another production mechanism for J/ψ, as shown pictorially in Figure 2.4.

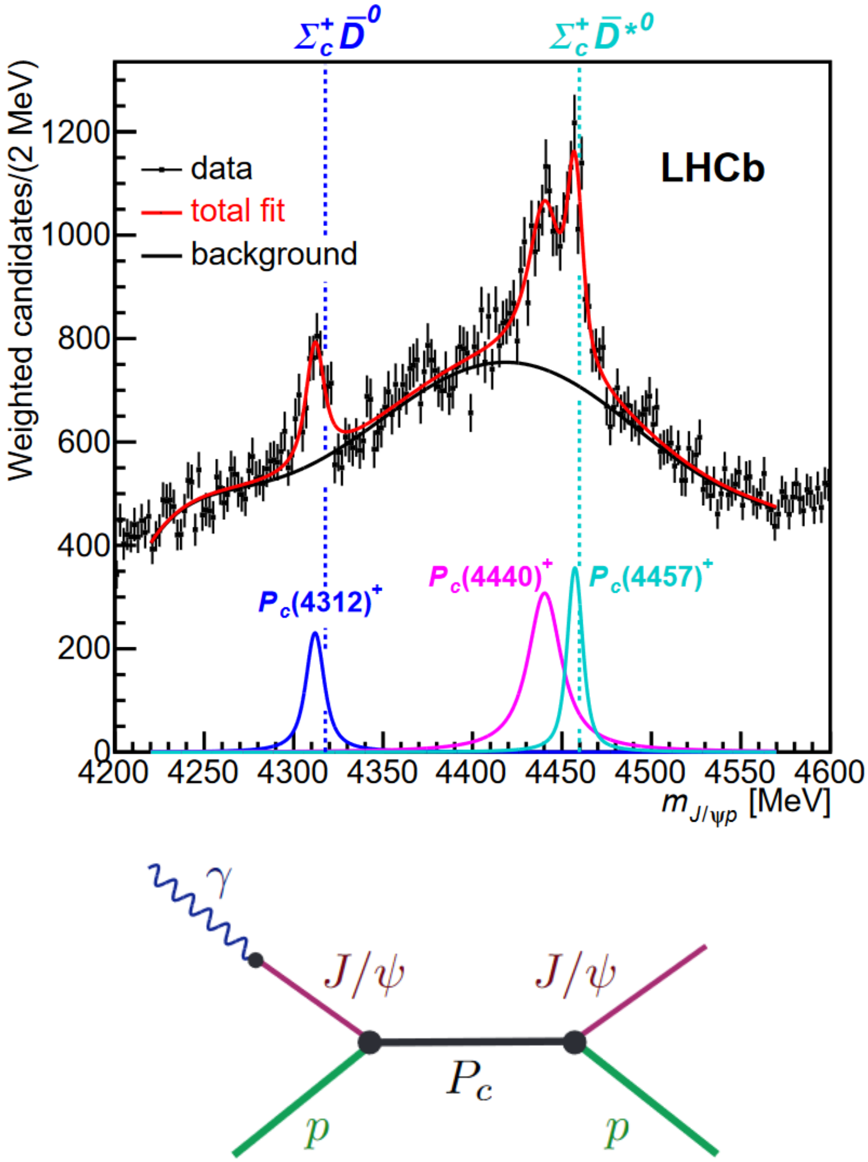


Figure 2.4: Top: P_c^+ pentaquark resonances in the $J/\psi p$ invariant mass spectrum [34]. The $\Sigma_c^+\overline{D}^0$ and $\Sigma_c^+\overline{D}^{*0}$ production thresholds are clearly indicated. Bottom: A diagram of P_c photoproduction in the VMD framework, taken from [29].

After the LHCb collaboration announced its discovery, the structure of these pentaquarks were modelled in several different ways. An initial claim was made that the signal observed by LHCb was due to kinematic effects [34]. These explanations of kinematic effects do not apply to photoproduction experiments [35] and a first motivation for measuring $J/\psi p$ in photoproduction experiments is therefore to confirm the results presented by the LHCb collaboration.

Several other models have also attempted to explain the structure of the P_c^+ states. The first of these models explains the P_c^+ states as composites of a heavy compact bound $c\bar{c}$ state with light nucleon quarks [36]. These are referred to as hadro-charmonium states. A second one interprets the P_c^+ states as hadronic molecules consisting of a charmed baryon and a charmed meson weakly coupled to one another [37]. This hypothesis is strongly supported by the proximity of the $P_c(4440)^+$ and $P_c(4457)^+$ to the $\Sigma_c^+\overline{D}^0$ and $\Sigma_c^+\overline{D}^{*0}$ production thresholds. A third interpretation describes the P_c^+ states as two tightly correlated di-quarks and an anti-quark [38]. The various models for the internal structure of the P_c^+ states are shown in Figure 2.5. Distinguishing between these interpretations of the P_c^+ states provides another motivation to study $J/\psi p$ photoproduction.

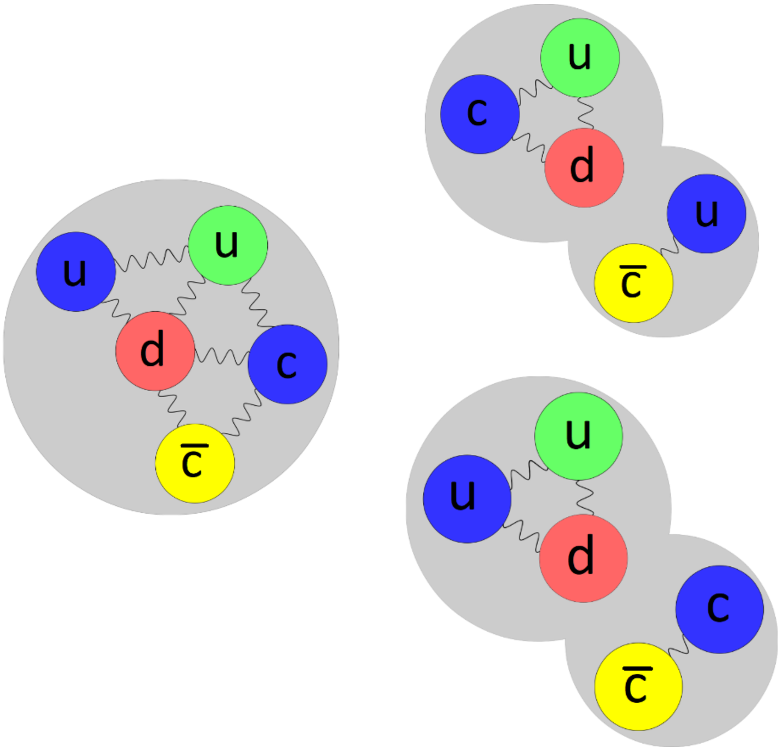


Figure 2.5: Pictorial representation of the P_c^+ models. The hadro-charmonium model is represented in the bottom right, with the molecular interpretation shown in the top right. The leftmost diagram represents two tightly correlated di-quarks and an anti-quark. Adapted from [39].

The first significant difference, notably between the hadro-charmonium and molecular interpretations, resides in their preferred decay modes. The hadro-charmonium models favour decays to charmonium states with hidden charm, such as J/ψ, and an ordinary baryon, such as the proton, with decays to open charm states being heavily suppressed [36]. The molecular model predicts the opposite, with the molecular states decaying predominantly to a charmed baryon and a charmed meson [37]. Precisely measuring the branching ratio of the decays to charmonium states and an ordinary baryon and the branching ratio to a charmed baryon and charmed meson would allow to distinguish between these two models. Although this would require a large collaborative effort between various experimental facilities, in the meantime at least placing an upper bound on the branching ratio in photoproduction experiments gives an indication towards the structure of these proposed P_c^+ pentaquark states.

2.2.2 Previous Measurements of the J/ψ Total Cross Section

The most direct way to study the J/ψ production mechanism is to measure the J/ψ total cross section as a function of photon energy. In fact, Brodsky *et al.* [27] originally sought to resolve tension between measurements of the total cross section of J/ψ photoproduction at SLAC [40] and Cornell [41]. The total cross section at Cornell seemed to indicate an unexpected increase in the total cross section at low photon energy, which was predicted to be due to the three-gluon exchange mechanism [27, 28]. However the lowest photon energy point, measured at Cornell, was situated at 11 GeV, whereas the main interest is to study the total cross section near the 8.2 GeV J/ψ photoproduction threshold.

More recent measurements taken by the Gluonic Excitation Experiment (GlueX) collaboration [42] are shown in black in Figure 2.6, along with the SLAC and Cornell measurements in red and blue respectively. The red dotted lines correspond to the theoretical predictions for the contribution of the two-gluon and three-gluon exchange to the total cross section and the incoherent sum of both. As shown the three-gluon exchange dominates closer to the threshold, with the two-gluon exchange dominating at higher energies. The black dotted line corresponds to a theoretical prediction for the total cross section which implies a large gluonic contribution to the nuclear mass [44], which will be discussed in more detail in the next section as it pertains to the nucleon gravitational form factors. The blue curve corresponds to the expected total cross section when taking into account the contribution from P_c^+ pentaquark production [43], assuming a branching fraction of the $P_c^+(4440)$ to $J/\psi p$ of 1.6%. The predictions are heavily model dependent and here rely on VMD based approaches. The GlueX measurements were not able to directly observe the P_c^+ pentaquarks, instead placing upper bounds on the branching fraction as 4.6%, 2.3%, and 3.8% for $P_c^+(4312)$, $P_c^+(4440)$, and $P_c^+(4457)$ respectively [42]. These upper limits on the branching fractions are an order of magnitude lower than the predictions in the hadro-charmonium scenario without excluding the molecular model [36, 45, 46].

As shown in Figure 2.6, the current measurements of the J/ψ photoproduction total cross section seemingly support the dominance of higher-twist contributions to the cross section at lower energies. The GlueX measurements do not favor either pure two- or three-hard-gluon exchange separately, and a combination of the two processes is required to fit the data adequately [42]. Here the two- and three-gluon exchanges are assumed to be non-interfering [42]. The previous SLAC measurements seemingly agree better with only the two-gluon exchange mechanism but do not significantly disagree with the incoherent sum of the higher-twist contributions. As such the two- and three-gluon exchange mechanisms are well supported by the available existing data for J/ψ photoproduction on the proton. Other two-gluon exchange models have also shown that they can suitably reproduce the J/ψ cross section data from threshold up to center of mass energies of 30 GeV [47]. No such data in the near-threshold region exists for J/ψ photoproduc-

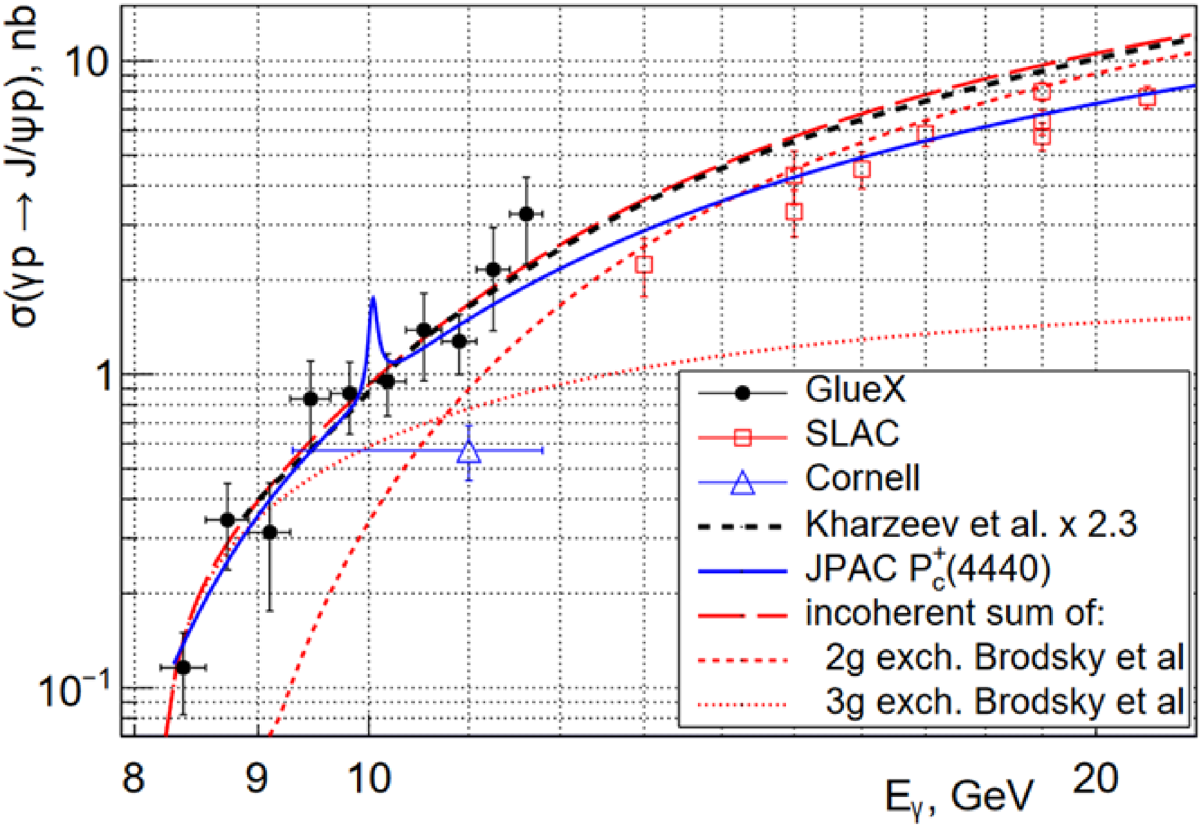


Figure 2.6: The near-threshold total cross section of J/ψ photoproduction on the proton as a function of photon beam energy from GlueX [42], SLAC [40] and Cornell [41] and compared to theoretical predictions [27,43,44]. Taken from [42].

tion on the neutron. Two-gluon exchange photoproduction is isospin invariant, and comparing the total cross section on the bound proton or bound neutron will enable to test the assumption that these production mechanisms are dominant close to threshold. Finally, note that GlueX has since produced higher statistics measurements of the J/ψ total cross section shown in Figure 2.7 which agree well with their previous measurements [48].

2.3 Probing the Mechanical Structure of the Nucleon

As alluded to in the introduction, J/ψ near-threshold photoproduction can be used to probe the gluonic content of the nucleon. For example, J/ψ photoproduction can be related to the distribution of gluons within the nucleon [49]. Based on the two-gluon exchange mechanism, a dipole parametrisation for the gluonic form factor which describes the distribution of gluons within the nucleon is written as:

$$F(t) \propto \left(1 - \frac{t}{m_g^2}\right)^{-2} \tag{2.8}$$

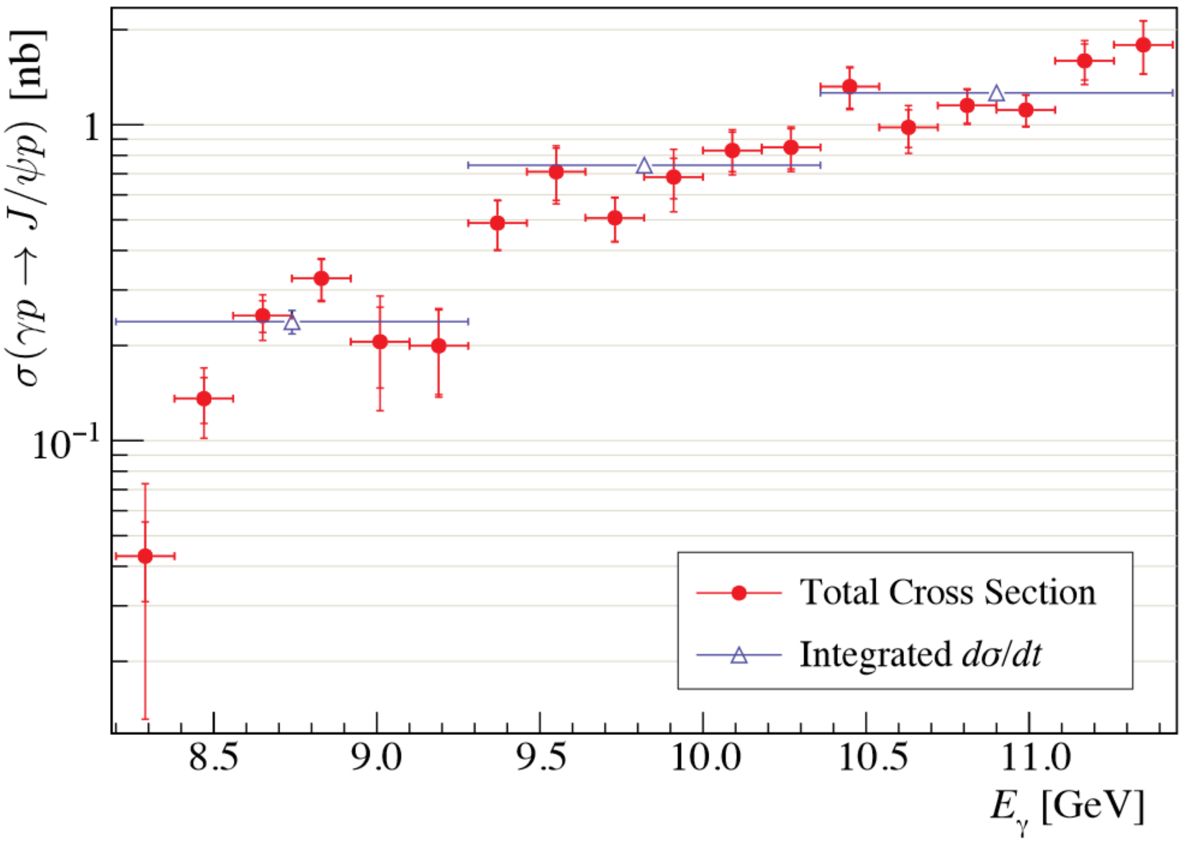


Figure 2.7: Most recent GlueX measurements of the J/ψ total cross section. Taken from [48].

where the mass scale parameter m_g^2 is assumed to be $m_g^2 \simeq 1$ GeV^2. The J/ψ photoproduction differential cross section as a function of t is then related to this form factor as:

$$\frac{d\sigma}{dt} = F(t)^2 \tag{2.9}$$

This calculation can then be compared to the SLAC [40] and Cornell [41] measurements of the differential cross section. As shown in Figure 2.8, there is good agreement between the prediction based on the dipole parametrisation and the measurements of the J/ψ differential cross section. The two-gluon exchange model was also shown to correctly reproduce the t dependency of the ϕ photoproduction differential cross section [49].

As mentioned previously, a different approach to predicting the J/ψ total cross section [44] implies a large gluonic contribution to the nuclear mass. Here, J/ψ photoproduction is first related to J/ψ-nucleon scattering based on VMD. The total J/ψ-nucleon cross section is then expressed in terms of nucleon gluonic form factors, with parametrisation [50, 51] which include measurements at HERA [52]. As shown in Figure 2.9, the real part of the J/ψ-nucleon scattering amplitude then dominates the cross section near-threshold, explaining the enhancement at low energy seen in the Cornell [41] measurements as discussed previously.

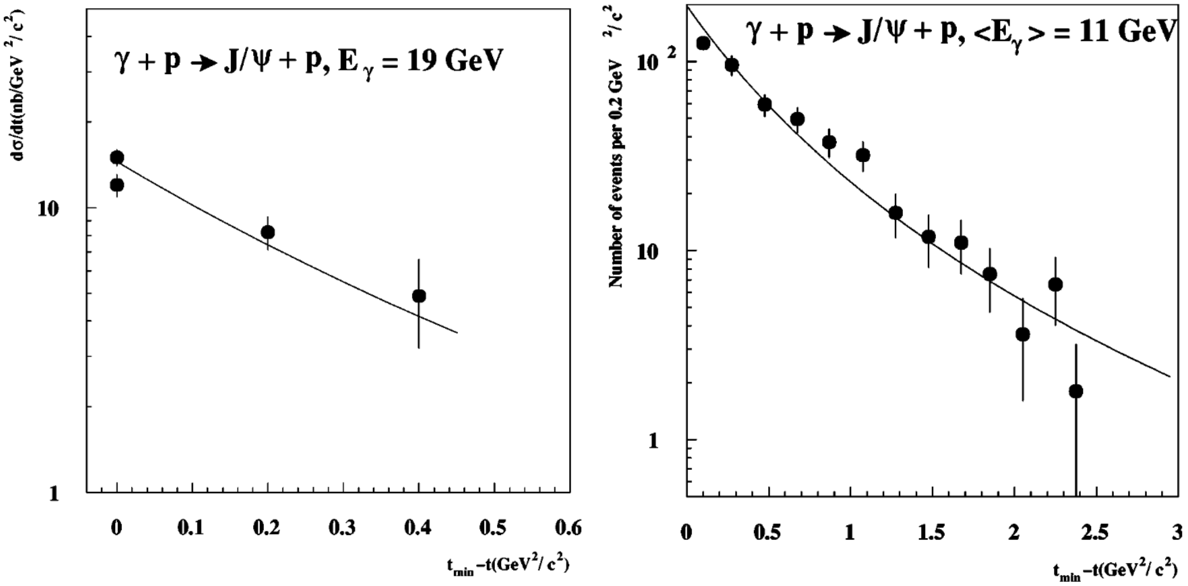


Figure 2.8: SLAC [40] (left) and Cornell [41] (right) measurements of the J/ψ differential cross section compared to the theoretical predictions based on a dipole form for the gluonic form factor. Taken from [49].

The real part of the J/ψ-nucleon scattering amplitude was also shown to contain the anomalous trace of the energy-momentum tensor [53]. The running coupling constant of Quantum Chromodynamics (QCD) is dependent on the mass scale, unlike classical field theory which is invariant under scale transformations. This manifests as a non-vanishing trace to the energy momentum tensor (EMT), named the trace anomaly [54], which is a key component of the origin of mass according to QCD. As explained in Chapter 1, the trace anomaly can be linked to the mass of the nucleon [8]. The above relation between J/ψ production and the trace anomaly of QCD was further expanded based on a holographic QCD approach to show how the trace anomaly impacts the shape of the total cross section and the differential cross section as a function t [16] in the near-threshold region. This then allows measurements of the cross section to be compared to theoretical predictions without the contribution of the trace anomaly [16]. A VMD based approach led to estimates of the b parameter which characterizes the magnitude of the anomaly contribution to the nucleon [55]. From the first GlueX measurements published in 2019 this parameter was estimated at $b = 0.07 \pm 0.17$ for the proton which suggests that the trace anomaly contribution to the proton mass is of order $\simeq 23.3\%$ [55]. This estimate was also modified to suggest the trace anomaly contribution to the proton mass ranges from 16 to 24% by including previous measurements of ϕ photoproduction [56].

Building upon the predicted relationship between the QCD trace anomaly and J/ψ photoproduction, the matrix elements of the EMT are directly related to J/ψ photoproduction [57, 58]. These elements are used to define the gravitational form factors [13] (GFFs) that encode mechanical properties such as the mass, pressure, and shear distributions of gluons within the nucleon,

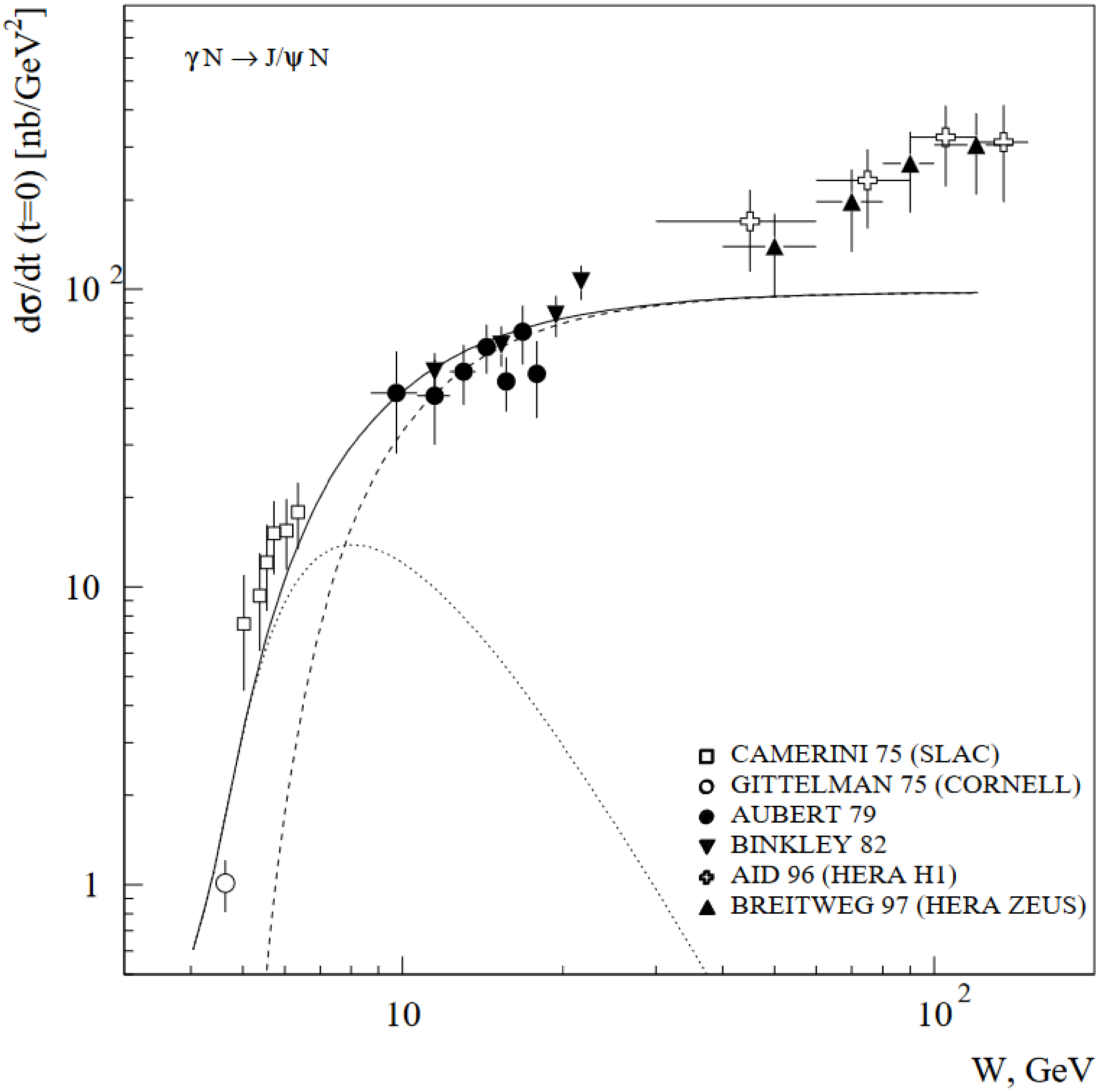


Figure 2.9: J/ψ photoproduction total-cross section data in solid points compared to theoretical predictions [44] with (solid line) and without (dashed line) the real part (dotted line) of the scattering amplitude. Taken from [44].

and notably the mass radius of the nucleon [57, 58]. Starting with the non-relativistic approximation for a weak gravitational field, it was shown that the mass distribution of a microscopic object with a weak gravitational field can be measured from its GFFs [57]. By rewriting the scalar gravitational form factor $G(q^2)$, for $-Q^2 = q^2$, as a linear combination of the other elements $G_i(q^2)$ of the EMT that fully describe the mechanical properties of the gluons within the nucleon, the mass radius of the nucleon is written as [57]:

$$< R_m^2 >= \frac{6}{M}\frac{dG}{dt}|_{t=0} \tag{2.10}$$

for M the mass of the nucleon. As mentioned previously, J/ψ near-threshold photoproduction is expected to by dominated by the $t-$channel exchange of gluons. It was then demonstrated that the scalar gravitational form factor can be measured from the differential cross section as a function of t of J/ψ near-threshold photoproduction [57]. Assuming a simple dipole form [57]:

$$G(t) = M(1-\frac{t}{m_s^2})^{-2} \tag{2.11}$$

for m_s a free parameter, the rms mass radius of the nucleon is then given by [57]:

$$< R_m^2 >= \frac{12}{m_s^2} \tag{2.12}$$

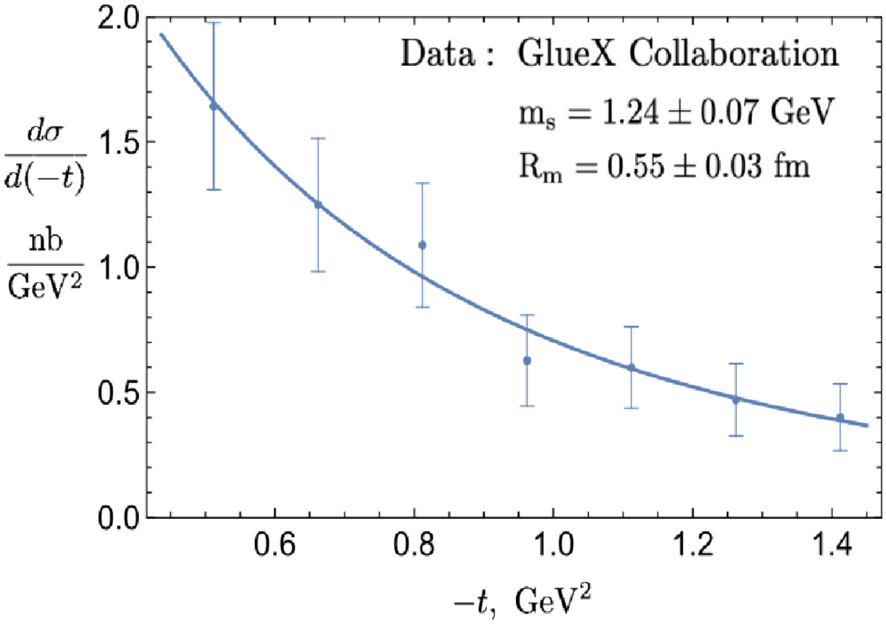


Figure 2.10: J/ψ photoproduction differential cross section as a function of –t. Data from the GlueX Collaboration [42], plot taken from [57].

Figure 2.10 shows the first measurement of the mass radius of the proton using GlueX measurements of the J/ψ differential cross section as a function of -t [42]. These constrain the free parameter m_s as $m_s = 1.24 \pm 0.07$ GeV giving a mass radius of $R_m = \sqrt{< R_m^2 >} = 0.55 \pm 0.03$ fm. Figure 2.11 shows more recent GlueX measurements of the J/ψ differential cross section measured in three E_γ bins ranging from 8.20-9.28 GeV, 9.28-10.36 GeV and 10.36-11.44 GeV. The mass radius measured in these bins are respectively 0.619 ± 0.094 fm, 0.464 ± 0.024 fm and 0.521 ± 0.020 fm [48]. In the lowest E_γ bin the differential cross section seems to increase at -t greater than three GeV^2. This could be an early indication that the assumption of the dominance of the two-gluon exchange is invalid in these kinematics, as will be discussed in the next section. The data points at larger -t were not used to measure the mass radius in this lowest E_γ bin. Other measurements made by the J/ψ-007 collaboration based in Hall-C at JLab are in good agreement as they estimate the mass radius to be $R_m = \sqrt{< R_m^2 >} = 0.52 \pm 0.03$ fm [58].

Lattice QCD predictions of the mass radius of the nucleon are also in good agreement with the GlueX and Hall C results [59], placing the mass radius at $R_m = \sqrt{<R_m^2>} = 0.517$ fm. Note that other calculations based on other models yield different results for the mass radius in [59], but more detail is given to explain why the prediction of $R_m = \sqrt{<R_m^2>} = 0.517$ fm is most reliable.

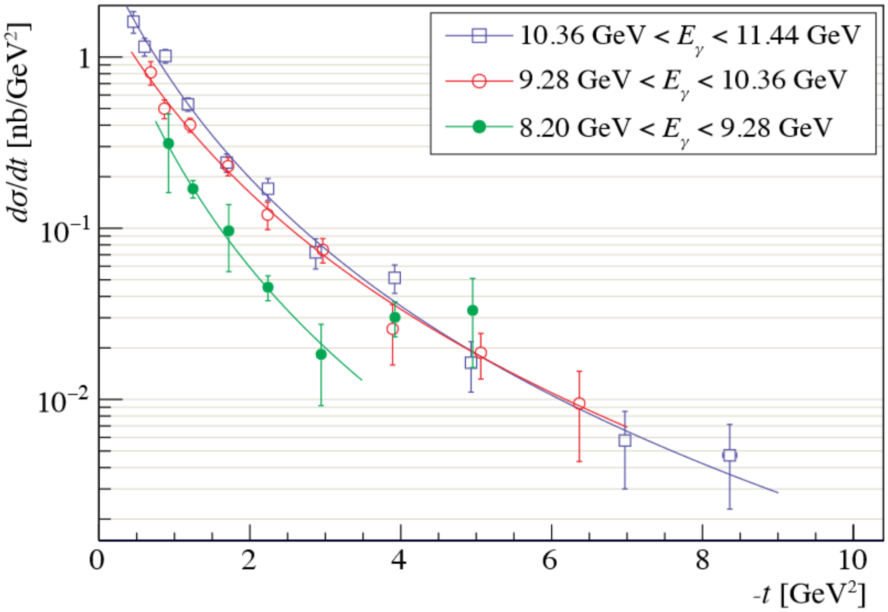


Figure 2.11: J/ψ photoproduction differential cross section as a function of –t. Taken from [48].

The quoted Hall C estimate of $R_m = \sqrt{<R_m^2>} = 0.52 \pm 0.03$ fm is averaged for photon energies above 9.7 GeV. Shown in Figure 2.12 is the mass radius estimated at Hall C in various photon energy bins. Good agreement is shown with GlueX and the lattice QCD predictions above 9.7 GeV. However, below this limit, the mass radius is estimated to be smaller on average [58]. This may be an indication that the three-gluon exchange does indeed dominate the two-gluon exchange closer to the J/ψ production threshold. In that case the mass radius cannot be estimated from J/ψ photoproduction.

The charge radius of the proton, measured at $R_C = 0.8409 \pm 0.0004$ fm [60], is statistically significantly larger than the mass radius of the proton. Although seemingly puzzling at first, it is important here to remember that the charge radius is extracted from the coupling of the photon to quarks, whereas the mass radius results from the coupling to gluons [57]. Simply put, it appears that the gluon radius of the proton is significantly smaller than its quark radius [57].

A QCD based analysis of near-threshold J/ψ production allowed to calculate the GFFs from the GlueX measurements [61]. The Hall C measurements allowed for a more precise determi-

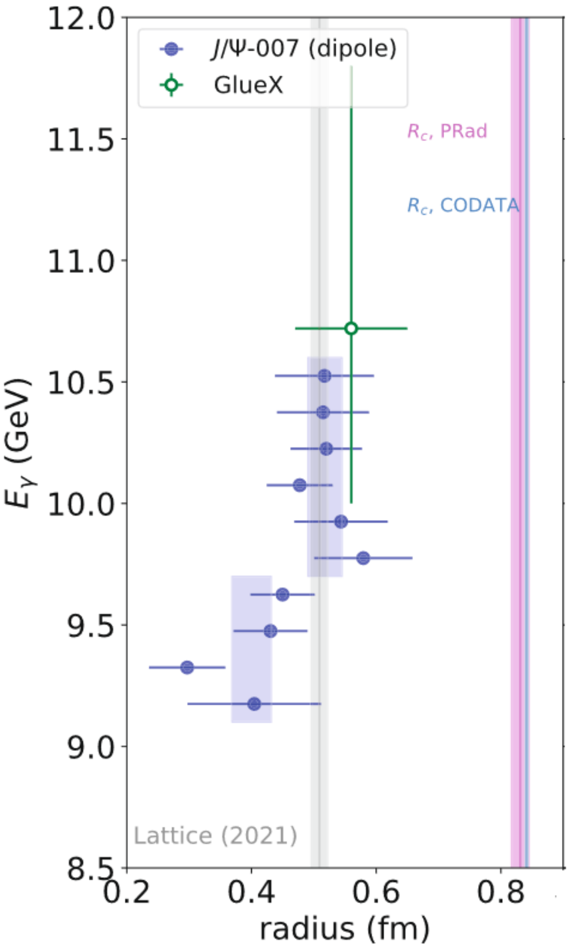


Figure 2.12: The mass radius of the proton measured at Hall C [58] for various bins in photon energy, compared to the first GlueX [57] measurement and a lattice QCD prediction [59]. Also shown is the charge radius of the proton from PRad and CODATA [60]. Taken from [58].

nation of the D_g GFF which encodes the shear forces and pressure distribution of the gluonic content of the nucleon, and the A_g GFF which is related to the momentum fraction of gluons in the nucleon [58]. The results are shown in Figure 2.13 when calculated with a holographic QCD approach in orange or with a GPD based approach in green [58]. The holographic QCD approach is in good agreement with the lattice QCD predictions [59].

The mass radius of the proton [62] and the bound proton and neutron in deuterium [63] has also been extracted from the photoproduction of light vector mesons ω and ϕ. For the free proton, the measurements on ϕ place the average mass radius at $R_m = 0.67 \pm 0.10$ fm [62] with those on ω giving $R_m = 0.68 \pm 0.03$ fm [62]. Figure 2.14 shows the mass radius as a function of the mass of the vector meson used to probe the gluonic contents of the proton. For the bound proton and neutron, the average mass radius extracted from ω photoproduction gives $R_m = 0.741 \pm 0.028$ fm and $R_m = 0.795 \pm 0.092$ fm for the proton and neutron respectively [63]. These results first off suggest consistency between the neutron and proton mass radius, within

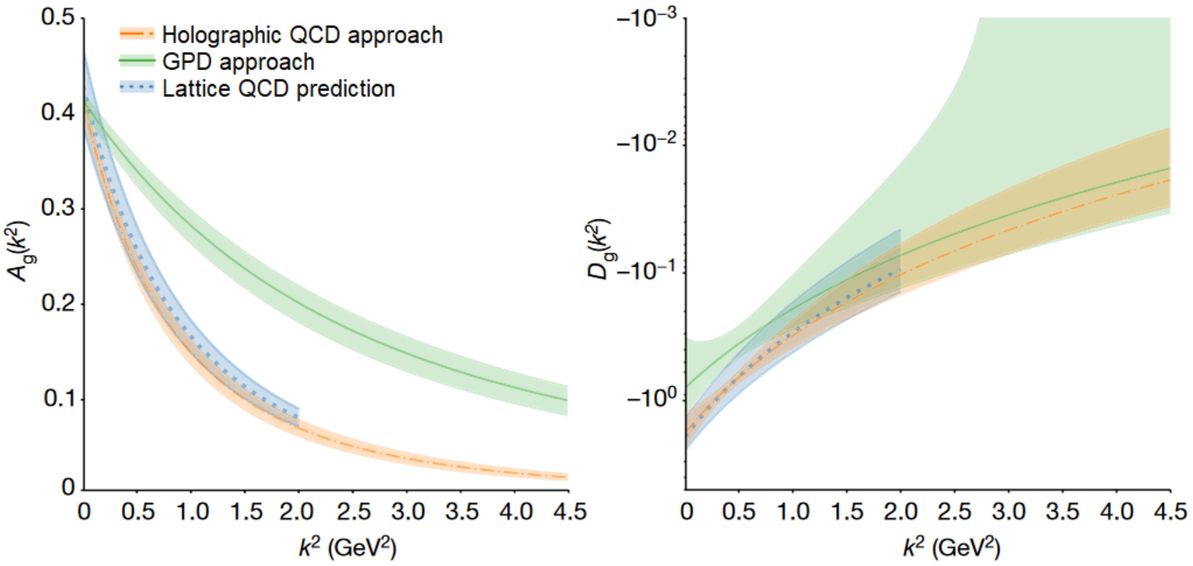


Figure 2.13: The A_g (left) and D_g (right) GFFs as a function of $k^2 \equiv |t|$. The GFFs are extracted using a holographic QCD approach (orange) or a GPD approach (green) and are compared to the lattice QCD predictions (blue). Taken from [58]

the measured statistical uncertainty. The difference between the bound proton and free proton was explained as potentially due to nucleon swelling in nuclear medium [63].

This concept of nucleon swelling was initially proposed as a solution to the EMC effect [64] which is the observation that the cross section for deep inelastic scattering (DIS) from an atomic nucleus is different from that expected for the same number of free protons and neutrons, essentially suggesting that the nuclear medium modifies parton distribution functions. This was initially demonstrated by the European Muon Collaboration when measuring a ratio significantly different from 1 in DIS from iron and deuterium targets [65], but in principle should apply to any nuclear medium. The suggestion is then that the quark constituents of the nucleon within a nucleus are shared with other nucleons [64, 66], leading to a de-confinement scenario whereby quarks can go between nucleons [66]. This would then affect the confinement size of quarks within the nucleon, with the effective quark radius about 15% bigger in the nucleus than in a free nucleon [66]. It was further suggested that measuring the charge radius of the nucleon within a nuclear medium would then allow to verify the nucleon swelling model [67] as the charge radius is intimately linked to the distribution of quarks within the nucleon. Measuring the mass radius of the nucleon within the nucleus would then allow to measure a potential swelling of the effective gluon radius.

The mass radius measurements obtained from light vector meson photoproduction bring tension to those measured with J/ψ photoproduction, as shown in Figure 2.14. It should be noted that the initial suggestion for measuring the nucleon mass radius relied on the production of

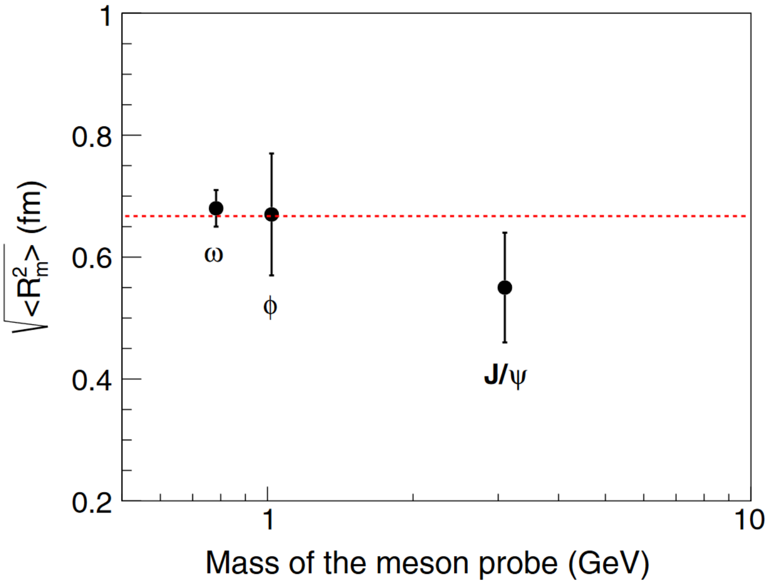


Figure 2.14: The mass radius of the proton measured from the photoproduction of ω, ϕ and J/ψ. The dashed line is set at the world average of $R_m = 0.67 \pm 0.03 fm$. Taken from [62].

heavy quarkonia [57] such as J/ψ and Υ which are composed of bound $c\bar{c}$ and $b\bar{b}$ quark antiquark pairs, whereas ϕ is composed of a $s\bar{s}$ pair and ω is a superposition of $u\bar{u}$ and $d\bar{d}$ pairs as $\frac{1}{\sqrt{2}}(u\bar{u} + d\bar{d})$. The reason for why these measurements were suggested for heavy quarkonia is due to the fact that their scattering amplitudes are dominated by the exchange of gluons and their main production mechanism can be modelled as a two-gluon exchange, as detailed previously. This is not the case for lighter vector mesons that are mainly produced via Pomeron exchange [68, 69]. The measurements of the mass radius from ω and ϕ production rely on the assumption that Pomeron exchange is a good approximation to two-gluon exchange [62, 68]. In particular it was shown that the gluonic form factor based on two-gluon exchange can adequately describe ϕ photoproduction [49]. The tension shown in Figure 2.14 might be an indication that these assumptions on the production mechanism are invalid [62]. Another suggestion is that the difference between the mass radius measured from J/ψ and that measured from the lighter vector mesons is due to a size or mass effect of the color dipole [62] which would suggest a dynamic response of the gluonic content of the proton when probed with different vector mesons.

2.4 Model Dependence

All the above theoretical work rely on certain assumed models. For example, the calculation of a gluonic form factor, the measurement of GFFs and of the nucleon mass radius all rely on the assumption of a two-gluon exchange being the dominant production mechanism for J/ψ photoproduction. However it was also suggested that the J/ψ near-threshold photoproduction could be dominated by open charm production [70]. The argument is as follows, in the strangeness sector, open strange photoproduction reactions such as $\gamma p \to K^+\Lambda$ or $\gamma p \to K^+\Sigma^0$ have a much larger cross section than that of near-threshold ϕ meson production, $\gamma p \to \phi p$ [70]. In this case the mass radius of the nucleon cannot be measured from ϕ photoproduction near-threshold as ϕ photoproduction is dominated by open strange photoproduction in this near-threshold region and is not dominated by two-gluon exchange. This might explain why the measurement of the mass radius from ϕ photoproduction is incompatible with those from J/ψ photoproduction. It is not unreasonable to suppose that open charm production such as $\gamma p \to \Lambda_c\overline{D}$ or $\gamma p \to \Lambda_c\overline{D}^*$ would be much larger than J/ψ photoproduction, in which case these would contribute significantly to J/ψ production [70].

Figure 2.15 shows the Feynman diagram for the open charm production of $J/\psi p$ where $\Lambda_c\overline{D}^{(*)}$ are first produced by the interaction of the photon with the proton target, before re-scattering into $J/\psi p$ [70]. The cross section for the associated process can then be calculated and is shown in Figure 2.16. Not only does the cross section appear to have the right order of magnitude, it also demonstrates a shape compatible with the data.

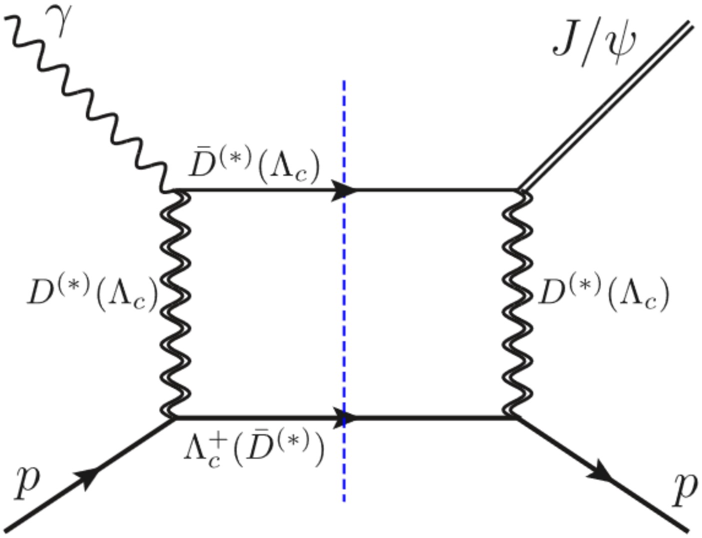


Figure 2.15: Feynman diagram for the open charm production of $J/\psi p$. Taken from [70].

The open charm production mechanism must therefore be ruled out before any claims can be made on measuring properties such as the mass radius of the nucleon. The most straightforward path to do so would be to rule out the distinct cusp like structure of the cross section from the open charm re-scattering to $J/\psi p$. These are due to the production thresholds of $\gamma p \to \Lambda_c\overline{D}$ and $\gamma p \to \Lambda_c\overline{D}^*$ as shown in Figure 2.16 [70]. One could also think of measuring the cross sections for the $\gamma p \to \Lambda_c\overline{D}^{(*)}$ channels and compare them to the theoretical estimates associated with the cross section for the open charm production of $J/\psi p$ [70].

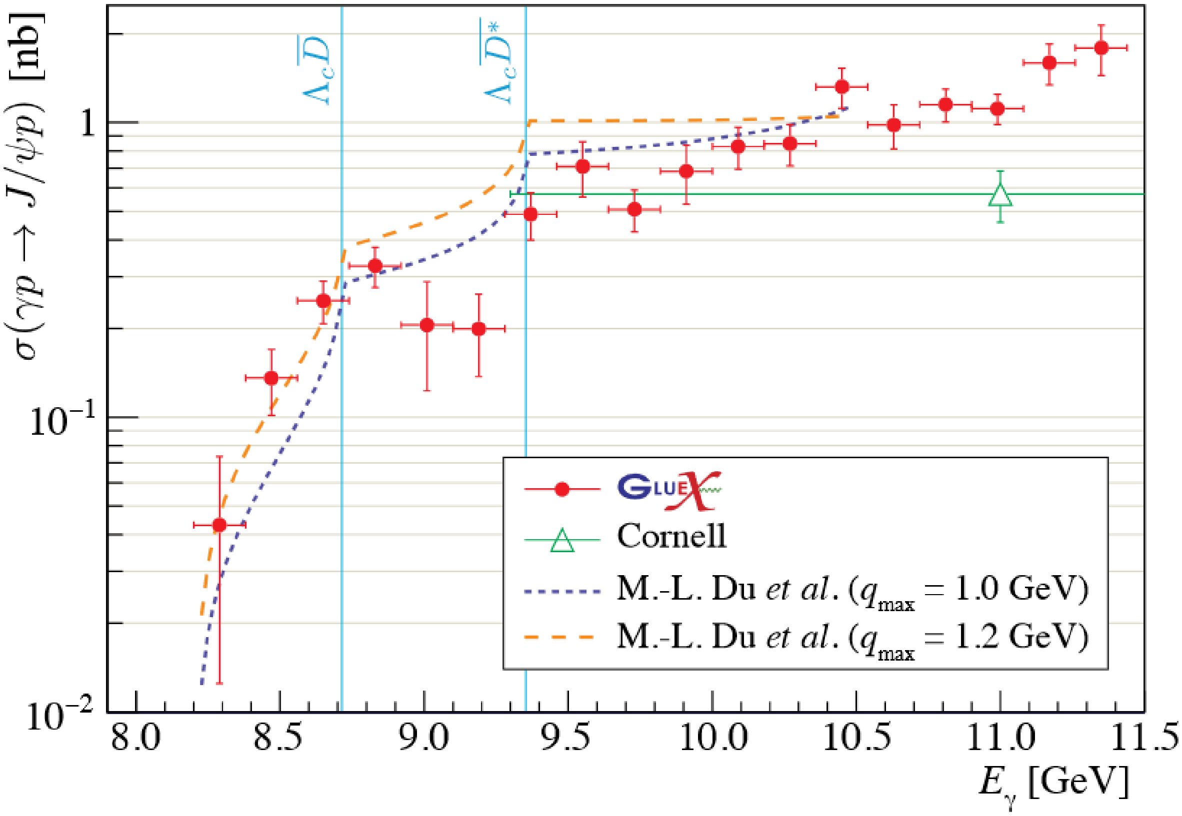


Figure 2.16: Prediction for the J/ψ total cross section due to the open charm production of $J/\psi p$ [70], which is consistent with the most recent GlueX measurements [48] in red. Here q_{max} refers to a threshold on Q^2 used in the flux calculations. The vertical dotted lines correspond to the $\Lambda_c\overline{D}^{(*)}$ production thresholds. Taken from [48].

Another important consideration is that of the J/ψ-nucleon scattering length. In the VMD framework, the scattering length $\alpha_{J/\psi p}$ can be related to the total cross section $\sigma_t(\gamma p \to J/\psi p)$ at threshold by [71]:

$$\sigma_t(\gamma p \to J/\psi p)|_{th} = \frac{q}{k}\frac{4\alpha_{em}\pi^2}{g^2_{J/\psi\gamma}}|\alpha_{J/\psi p}|^2 \qquad (2.13)$$

where k is the center of mass momentum of the incident photon at the $\gamma p \to J/\psi p$ threshold, α_{em} is the fine-structure constant, and $g^2_{J/\psi\gamma} = 5.58 \pm 0.07$ is the $\gamma \to J/\psi$ coupling, as determined from the $J/\psi \to e^+e^-$ decay width [71]. The $J/\psi p$ scattering length was estimated by extrapolating back to threshold the first measurements at GlueX of the total [71] and differential [72] cross sections, with the estimates placing the scattering length on a range of 0.003–0.025 fm. Another interesting results showed in Figure 2.17 hints to a nonlinear exponential increase of $\alpha_{Vp} \propto e^{\frac{1}{m_V}}$ with increasing $\frac{1}{m_V}$ [73]. Measuring the cross section of ϕ and ω photoproduction alongside that of J/ψ photoproduction would then provide an interesting way to test the scattering length estimates.

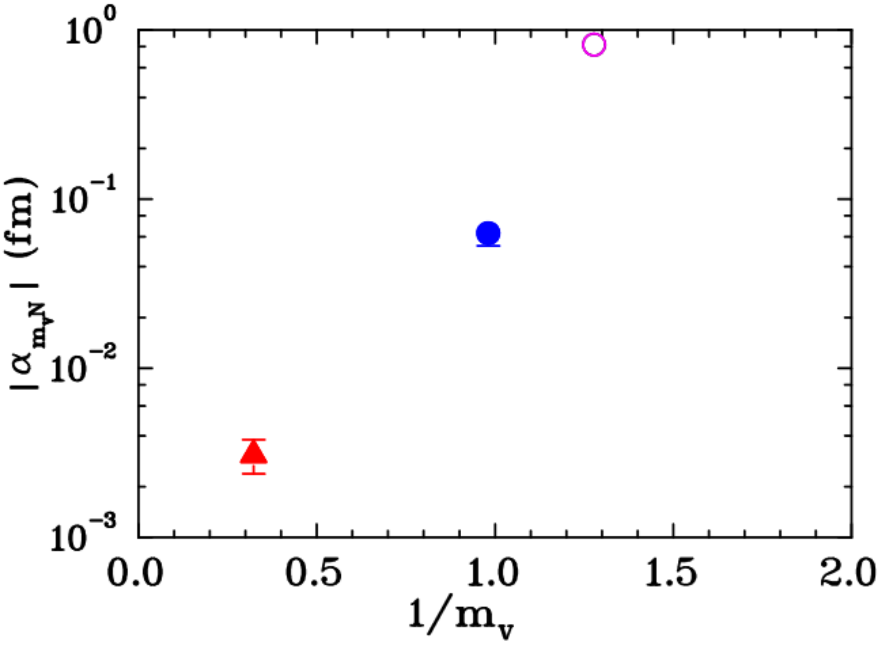


Figure 2.17: Estimates of the scattering length of ωp (magenta), ϕp (blue) and $J/\psi p$ (red) as a function of the inverse mass of the vector meson m_V. Taken from [73].

However these estimates of the $J/\psi p$ scattering length do not agree with previous theoretical predictions [71], and in particular that based on a two-gluon exchange mechanism which predicts a sizeable scattering length of order 0.25 fm [74], which is at least an order of magnitude more than what was calculated from GlueX data. The open charm model predicts the $J/\psi p$ scattering length to be on a range of 0.0002 to 0.003 fm [70], which is closer to the scale calculated from GlueX data. Additionally, the fact that the scattering length is much smaller than the typical size of a hadron [71] would suggest that the proton is transparent to J/ψ. This is in line with lattice QCD predictions [75] that suggest non-interacting charmonium and nucleon. This would in turn suggest that a strong coupling between the nucleon J/ψ channel with other two-hadron channels might be responsible for the existence of the P_c resonances [75]. Note that the most recent GlueX data was used to estimate the $J/\psi p$ scattering length at $(21.3 \pm 8.2)10^{-3}$ fm [48].

Another way to study the J/ψ production mechanism is to measure the J/ψ spin density matrix elements (SDMEs) [76]. These fully describe the angular distribution of meson production and decay, and can shed light on the production mechanism. For example, by using a linearly polarised photon beam, natural and unnatural parity exchange production mechanisms can be distinguished [77] by combining several SDMEs. Figure 2.18 shows the expected naturality for two-gluon or three-gluon exchanges, with the three-gluon exchange violating charge parity [78]. At present it is not clear if the open charm mechanism violates parity, but this could prove a suitable way to study the production mechanism of J/ψ. This would also allow to study the energy ranges at which two-gluon exchange dominates three-gluon exchange, and vice versa, establishing an energy regime where the measurement of the gluonic gravitational form factors is valid.

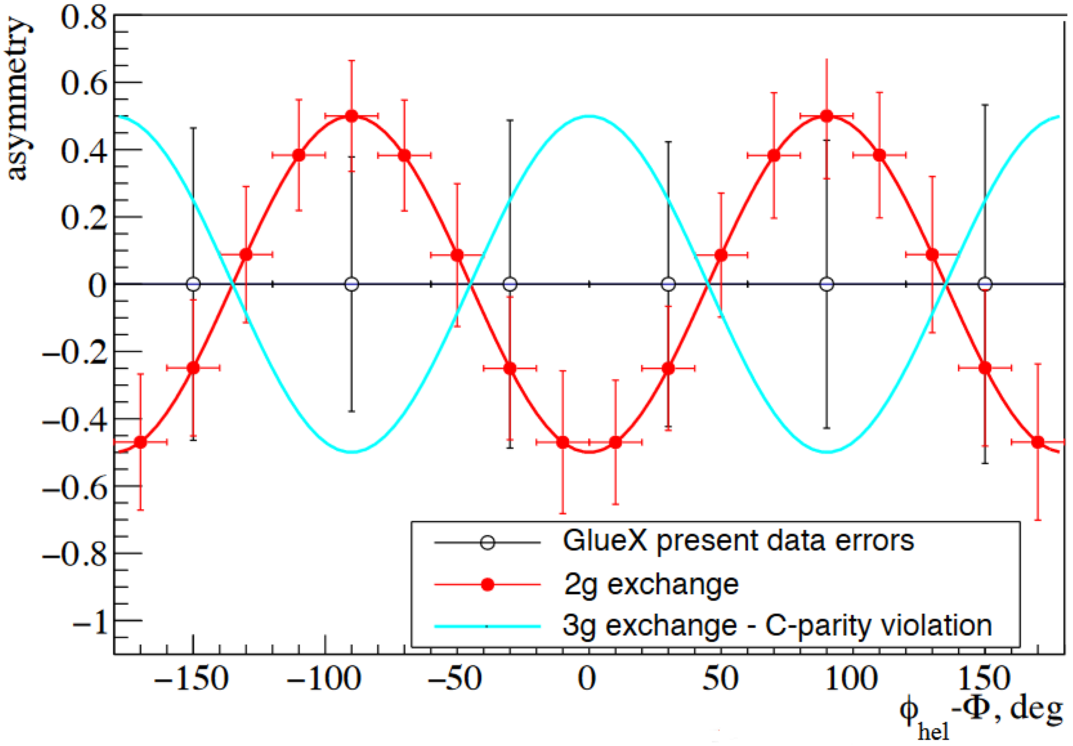


Figure 2.18: Expected naturality for the 2-gluon or 3-gluon exchange J/ψ production mechanism. Also shown are the most recent results from GlueX, hinting to the need for additional data. Taken from [78].

Another feature of Figure 2.18 is the current measurements of the J/ψ naturality at GlueX. Due to low statistics, these cannot yet distinguish between the natural and unnatural production mechanisms. Similarly, additional data would be required in order to identify or rule out the presence of the cusp like structure of the cross section due to open charm production. Potential beam energy or luminosity upgrades at JLab could allow for these measurements to be made [78]. Additionally the future Solenoidal Large Intensity Device (SoLID) [79] experiment at JLab's Hall-A will provide high statistics for J/ψ electroproduction. Unfortunately these experiments will not happen in the near future as they are yet to be commissioned, and as such distinguishing between open charm and two or three-gluon exchange in J/ψ photoproduction is as of yet unattainable.

Finally it must be pointed out that most of the above theoretical work relies on the assumption that VMD suitably describes J/ψ photoproduction. However, there have been some suggestions that VMD fails to suitably describe heavy vector-meson production [80] such as J/ψ as it cannot reliably estimate a photon to heavy vector meson transition strength or the momentum dependence of terms necessary to calculate the different reaction amplitudes [80]. Older theoretical work also claimed a strong energy dependence of the near-threshold J/ψ cross section due to a large transfer momenta in J/ψ photoproduction, which would invalidate VMD based models [71, 81]. Whilst these claims must be investigated and resolved by the theoretical hadronic physics community, one could point to the successes of VMD based models in describing the total and differential J/ψ cross sections as motivation enough not to discard VMD just yet. Furthermore, some approaches based on QCD [61] or holographic QCD [16] which do not rely on VMD have also related J/ψ photoproduction to the trace anomaly of QCD and GFFs.

2.5 J/ψ Quasi-real Photoproduction on the Proton and Neutron at CLAS12

As shown in the previous sections, J/ψ photoproduction offers a wealth of information on the structure of nucleons. This is especially relevant given that J/ψ can be used to probe the gluonic contents of the nucleon, which is typically hard to do in photoproduction experiments given that electrons or photons will not directly scatter off of the gluon. Several experiment proposals were therefore submitted to study J/ψ photoproduction at the CLAS12 detector located in JLab's Hall-B [29, 82, 83]. The previous sections introduced previous measurements of J/ψ photoproduction on the proton; CLAS12 can provide additional, independent measurements of J/ψ photoproduction on a liquid hydrogen (LH_2) target [29, 82], which essentially can be thought of as a free proton. However, given that CLAS12 also took data with a liquid deuteron (LD_2) target, the measurements of J/ψ photoproduction can also be made on the quasi-free bound proton and bound neutron within the deuteron, so called incoherent J/ψ photoproduction on the deuteron [83]. This would amount to a first measurement of near-threshold J/ψ photoproduction on the bound nucleon.

J/ψ photoproduction is expected to be isospin invariant due to the dominance of two- or three-gluon exchange as a production mechanism [83]. The expectation is therefore that the total and differential cross section of J/ψ photoproduction off the proton and neutron will behave similarly [83]. Nevertheless this will not only constitute a first measurement of J/ψ photoproduction off the bound neutron but also off the bound proton, and as such allows to measure and compare the bound proton gluonic form factors and mass radii to the free proton gluonic form factors and mass radii. As mentioned previously this could also provide some hints towards nucleon swelling within the nuclear medium, as significant differences in the mass radii of the bound and free proton would suggest a potential swelling of the effective gluon radius. Finally, J/ψ photoproduction off the bound proton and neutron allows for a direct comparison of J/ψ photoproduction on the proton and neutron.

The positively charged charmonium P_c^+ pentaquark candidates detected by the LHCb should have isospin neutral partners [83] that should be found in J/ψ production on the neutron. The detection of such states would add weight to confirm the existence of P_c^+ pentaquark states. Unfortunately the limited statistics available will not allow for a direct observation of the P_c^+ pentaquark states and their isospin partners. This will therefore not be the focus of the work presented in this thesis, but deserves to be mentioned here given that this body of work will pave the way for future experiments aiming to observe the isospin partners of the P_c^+ pentaquarks at CLAS12, assuming more data is taken towards these measurements.

The measurements presented in this thesis will be made in the following reaction channels:

$$ep \to e'J/\psi p \to (e')l^+l^-p$$

$$ep_{bound} \to e'J/\psi p \to (e')l^+l^-p$$

$$en_{bound} \to e'J/\psi n \to (e')l^+l^-n$$

where $l^{\pm}$ denotes as before a lepton (either $e^{\pm}$ or $\mu^{\pm}$) of positive or negative charge. Note that the the J/ψ branching ratio to an electron positron pair is equal to that to a di-muon pair. Given that CLAS12 operates with an electron beam it is only natural that it would be able to correctly detect and identify electrons. As such the most straightforward channel to investigate at CLAS12 is that of J/ψ produced on a hydrogen target, decaying to an electron and a positron. This channel will serve as a comparison to the results presented in this thesis but will not be discussed in detail as it has been studied by CLAS12 collaborators [84]. As mentioned above, J/ψ can also decay to an oppositely charged di-muon pair. Studying J/ψ production on a hydrogen target in its decay to a di-muon pair presents a very direct comparison to the measurement with J/ψ decaying to a di-electron pair. The corresponding total and differential cross sections for the di-muon decay channel were therefore measured as a part of the work presented in this thesis. Finally, the main objective of this thesis is to compare the photoproduction of J/ψ on the proton and neutron. The total and differential cross sections for incoherent J/ψ photoproduction on the proton and neutron will therefore be measured and compared in this thesis in the J/ψ decay to a di-electron pair. This will correspond to a first time measurement of J/ψ photoproduction on the bound neutron and bound proton.

It should also be noted in the above channels that the final states proposed here would only measure the two leptons and recoil proton, without measuring the scattered electron; the parentheses around e’ are there to denote that the scattered electron goes undetected. This is the so called untagged analysis [29]. In principle a similar tagged analysis could also measure J/ψ photoproduction at CLAS12 [29], notably by considering the missing mass of the electron and recoil nucleon which should have a peak at the J/ψ mass. This would have the additional advantage of measuring J/ψ photoproduction in both its decay to muons and electrons simultaneously. However, studies of the effective luminosity in both the untagged and tagged cases show that the effective luminosity is essentially decreased by an order of magnitude in the tagged analysis relative to the untagged analysis [29], as shown in Figure 2.19, which then leads to a corresponding decrease by an order of magnitude in the number of J/ψ that can be measured in the tagged analysis. This is due to the requirement of detecting a very forward going scattered electron that essentially will be very close to the beam line. Placing a detector so close to the beamline is complicated as it would be overloaded by the electron beam. This restricted angular coverage available for the forward going scattered electron then decreases the effective luminosity of the tagged analysis by an order of magnitude [29].

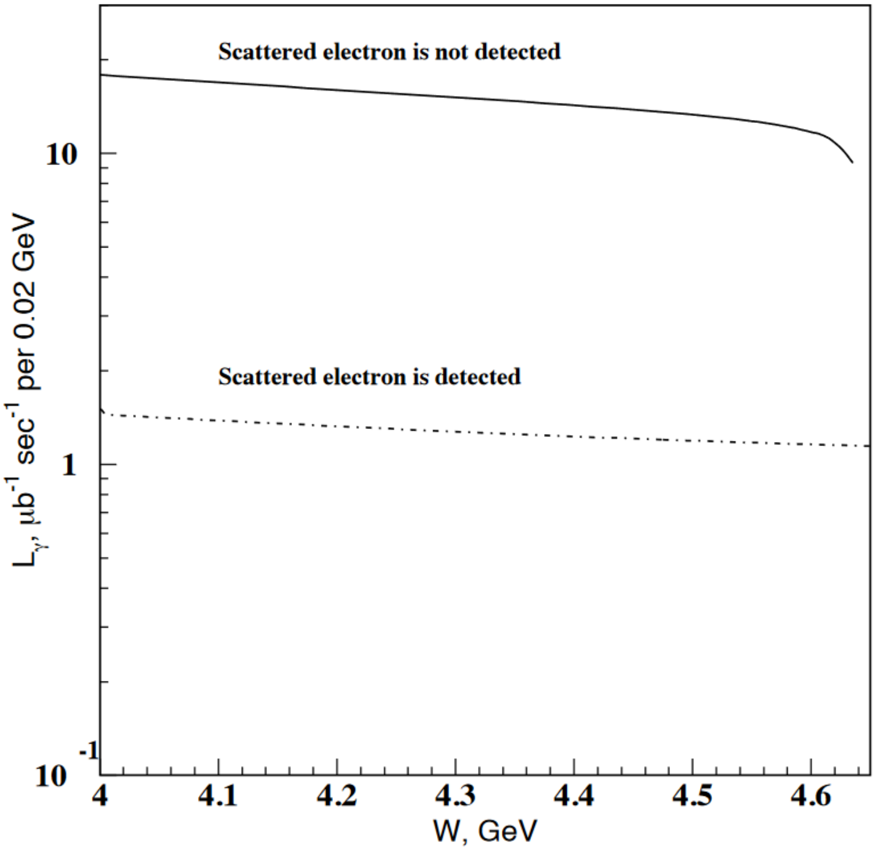


Figure 2.19: Comparison of the effective luminosities for the untagged and tagged analyses [29] over the center of mass energy range for J/ψ production.

Note that another channel could also be explored where J/ψ is produced on the deuteron itself, so called coherent J/ψ production on the deuteron [83]. This final state is distinguished from that of J/ψ produced on the bound proton or neutron by having a deuteron in the final state. A preliminary analysis investigated the feasibility of such a measurement. The conclusion was that due to a slightly different physics reach and the low number of J/ψ produced on the deuteron, priority would be given to the analyses mentioned above.

Chapter 3

The CLAS12 Detector

The aim of this chapter is to introduce the experimental apparatus used for the measurements of J/ψ photoproduction at CLAS12. This chapter starts by giving an overview of the CLAS12 detector, before focusing in on the most relevant subsystems in the context of J/ψ photoproduction. This chapter then ends with a description of the reconstruction framework that parses information from various detector subsystems into the high level information needed for the measurement of J/ψ photoproduction.

3.1 Overview

The Continuous Electron Beam Accelerator Facility (CEBAF) [85, 86] produces and delivers an electron beam to the different halls at JLab, with the CEBAF Large Acceptance Spectrometer (CLAS12) [88] located in Hall B. Electrons are produced by firing a laser beam onto a photocathode to release electrons via the photo-electric effect. The electrons are then drawn out with an anode and enter the CEBAF where they will be accelerated to the desired beam energy.

A schema of CEBAF is found in Figure 3.1. Electrons first travel through the linear accelerators arms where they acquire 1200 MeV from each passage [89]. Dipole magnets bend the beam through recirculating arcs as it loops through the linear accelerators four and a half times before being redistributed to the different halls once it has reached its desired energy. The beam is also split up into bunches every 2ns [89].

Once it enters Hall B, the electron beam is directed to a target of choice where an interaction between the electrons and the target produces a variety of final state particles which are then detected by the CLAS12 detector. CLAS12, pictured in Figure 3.2, was built to have full azimuthal angular coverage and a large acceptance in polar angle, allowing measurements to be made over large kinematic ranges [90]. Very low Q^2 coverage is enabled by the forward tagger (FT), whilst the forward detector (FD) covers the range of polar angles from 5 to 35 degrees,

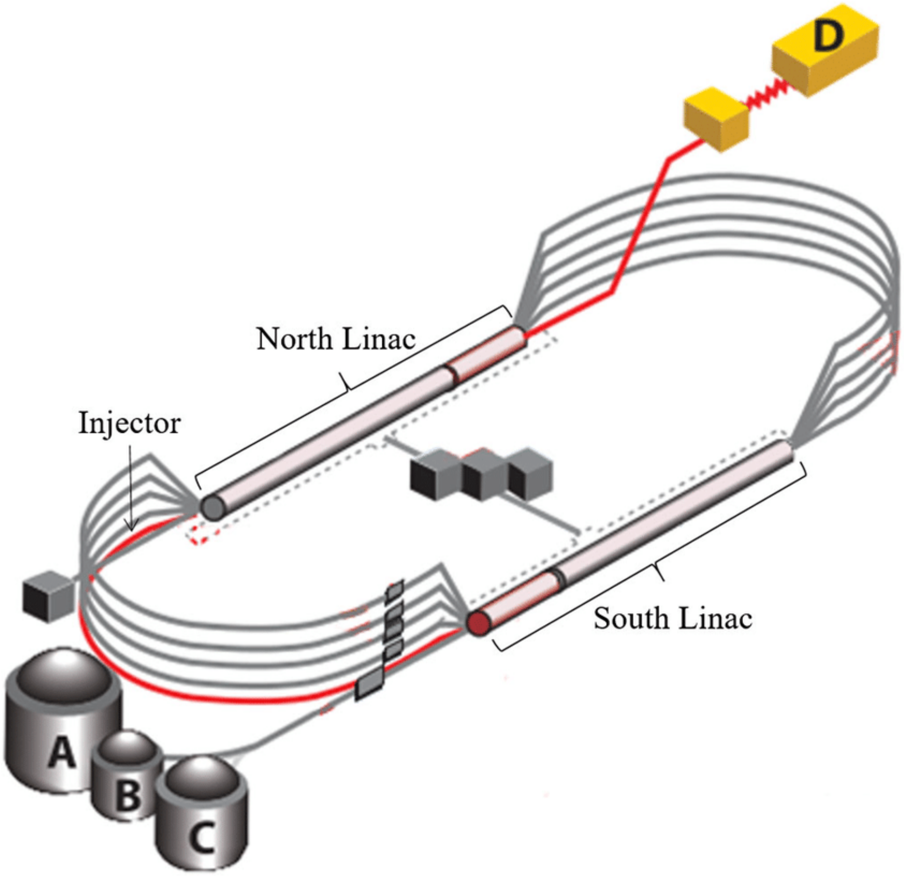


Figure 3.1: A diagram of CEBAF and the four experimental halls at JLab. Taken from [87].

with the forward detector further segmented into six sectors of azimuthal angle. The central detector (CD) then covers the range of polar angles from 35 to 125 degrees [88].

The FT was mainly designed to detect very forward going photons and electrons after they have interacted with the target, named scattered electrons. The FT is made of three constituents, the first being the FT electromagnetic calorimeter designed to measure the energy [89, 91] of the electrons or photons. The second constituent is a hodoscope that allows to separate electrons and photons that produce indistinguishable electromagnetic showers in the calorimeter [89, 91]. An electron is therefore identified by a hit in the hodoscope that is correlated in position and time with a hit in the calorimeter. Finally, the FT tracker measures the scattering azimuthal and polar angles of the electrons [89, 91]. The 3-momentum of the scattered electrons are then reconstructed from these scattering angles.

The FD measures and identifies particles scattered between 5 and 35 polar degrees in the full azimuthal range with 6 identical magnetic spectrometers that are split into several constituents. A diagram of the FD and its constituents is found in Figure 3.3. The High and Low Threshold Cherenkov Counter (HTCC and LTCC respectively) are used to resolve electrons from pions below 5 GeV and resolve pions from kaons [92, 93]. The Forward Drift Chambers (DC) use a

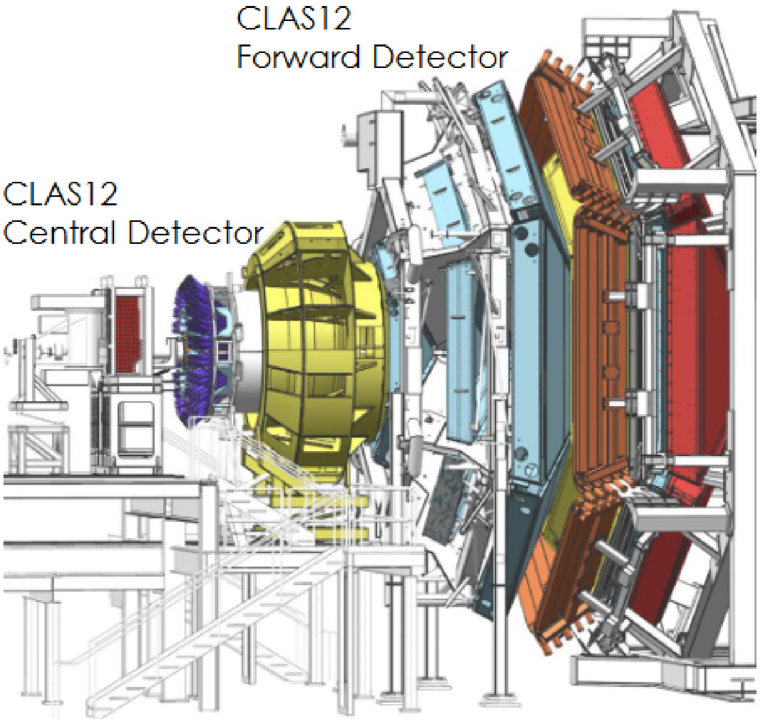


Figure 3.2: A diagram of the CLAS12 detector showing the central and forward detectors. Three of the forward detector's six sectors are visible. The forward tagger covering low polar angles is here hidden from view and is situated close to the beam-line. Taken from [88].

toroidal magnet to bend and record particle tracks through the drift chambers [94]. These tracks are then used to determine the charge of the particles and measure their momentum. The Forward Time Of Flight (FTOF) scintillators measure with high accuracy the velocity β of charged particles in the FD [95]. The FTOF was designed with hadron identification in mind, mainly to resolve pions from kaons and protons. Finally, the Forward Electromagnetic Calorimeters (ECAL) are used to identify electrons with momentum above 0.5 GeV, measure photons with momentum above 0.2 GeV and detect neutrons [96]. The ECAL is made of three separate calorimeters, a pre-shower calorimeter (PCAL), an inner and an outer calorimeter (ECIN and ECOUT respectively) to be able to absorb energy from electrons and photons with momentum above 5 GeV.

The CD was built to detect charged hadrons in the full azimuthal range, between 35 to 125 polar degrees and with momenta below 1.5 GeV [89]. A diagram of the CD and its constituents is found in Figure 3.4. The Central Tracking system, much like that in the FD, measures the momenta and charge of particles in the CD [98, 99]. The Central Time Of Flight scintillator counters measure the velocity of particles in the CD and are designed to resolve pions from kaons and pions from protons [100]. Finally, the Central Neutron Detector also made of scintillator

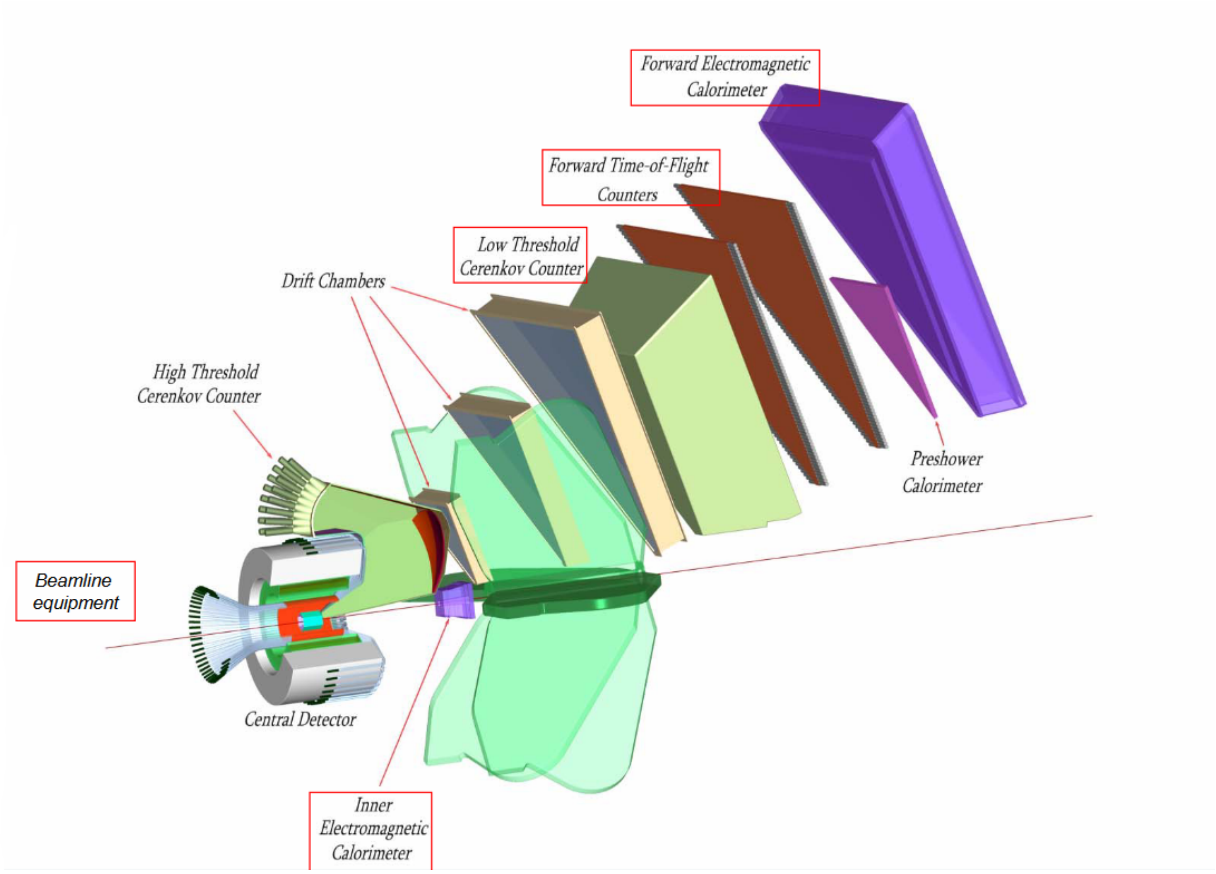


Figure 3.3: A diagram of the CLAS12 detector with an emphasis on the Forward Detector and its constituents. This plot was taken from [97].

counters was designed to detect neutrons [101].

A summary of the CLAS12 performance is shown in Table 3.1, demonstrating excellent performance in measuring the four momentum of both charged and neutral particles, along with good particle identification for charged particles.

3.2 The CLAS12 Forward Detector

The analysis of J/ψ photo-production at CLAS12 requires a di-lepton pair and recoil nucleon detected in the forward detector, with the scattered electron going undetected. As such it is worth spending a bit more time understanding the forward detector and its various subsystems so as to better understand the analysis of J/ψ photo-production. As discussed in the previous sections, the forward detector is composed of the ECAL, the FTOF, the HTCC and LTCC, and the DC.

The ECAL is composed of two separate subsystems, the electromagnetic calorimeter (EC) and the pre-shower calorimeter (PCAL), both composed of three views (U/V/W), as exemplified in 3.5. The EC itself contains two calorimeters, the EC inner (ECIN) and EC outer (ECOUT),

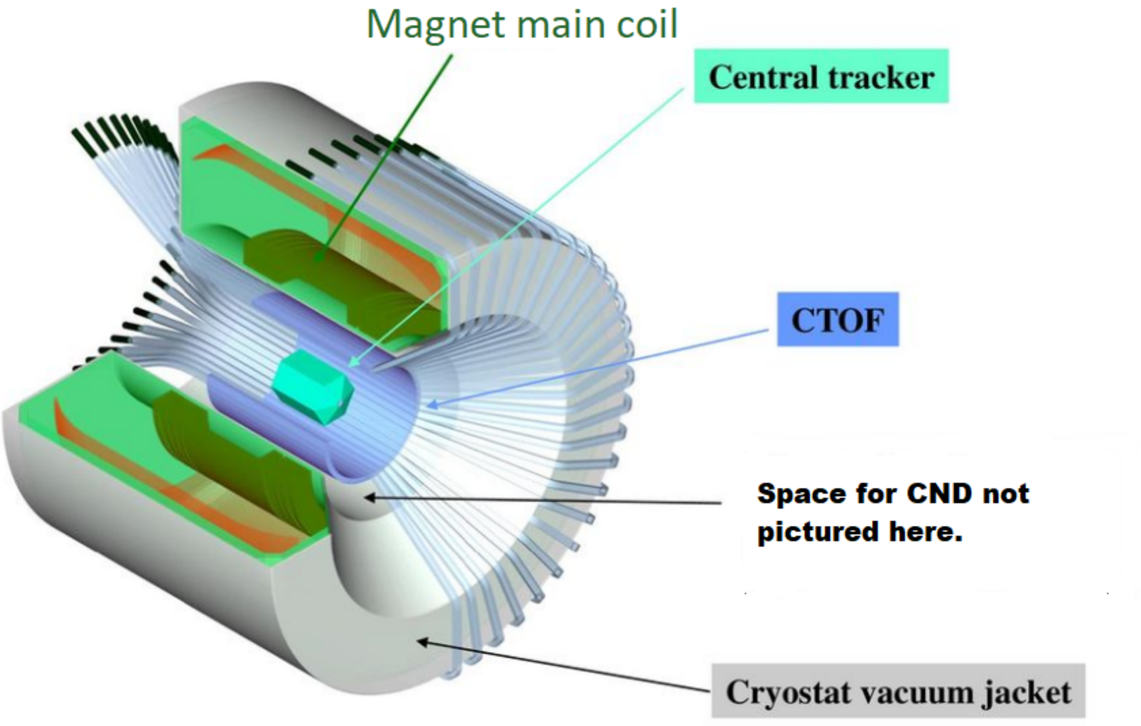


Figure 3.4: A diagram of the Central Detector and its constituents. Adapted from [102].

Parameter	Forward Detector	Central Detector
Charged tracks:		
Polar angular range (θ)	5°to 35°	35°to 125°
Polar angular resolution ($\delta\theta$)	<1 mr	<10 mr to 20 mr
Azimuthal angular resolution ($\delta\phi$)	<4 mr	<5 mr
Momentum resolution ($\frac{\delta P}{P}$)	<1% at 5 GeV	<5% at 1.5 GeV
Neutral particles:		
Polar angular range (θ)	5°to 40°	40°to 125°
Polar angular resolution ($\delta\theta$)	<4 mr	<10 mr
Energy/momentum resolution	$<0.1/\sqrt(E)$	<5%
Particle Identification e/π	0-4.5 GeV	N/A
π/p	Full Momentum Range	<1.25 GeV
K/π	<3 GeV	<0.65 GeV
K/p	<4 GeV	<1 GeV

Table 3.1: A summary of the CLAS12 detector performance adapted from [89].

which have 36 strips in each of their three views. The PCAL has 68 strips in the U view and 62 in V and W [96]. Particles impacting the calorimeter deposit energy, producing so called clusters. A hit in the calorimeters is defined as a strip having energy above a user defined threshold. Contiguous strips with hits are then grouped into one or more peaks, as each peak can be further subdivided based on the hit energy profile of the groupings. Finally, clusters are identified at the spatial intersection of a peak in each of the U/V/W views [96].

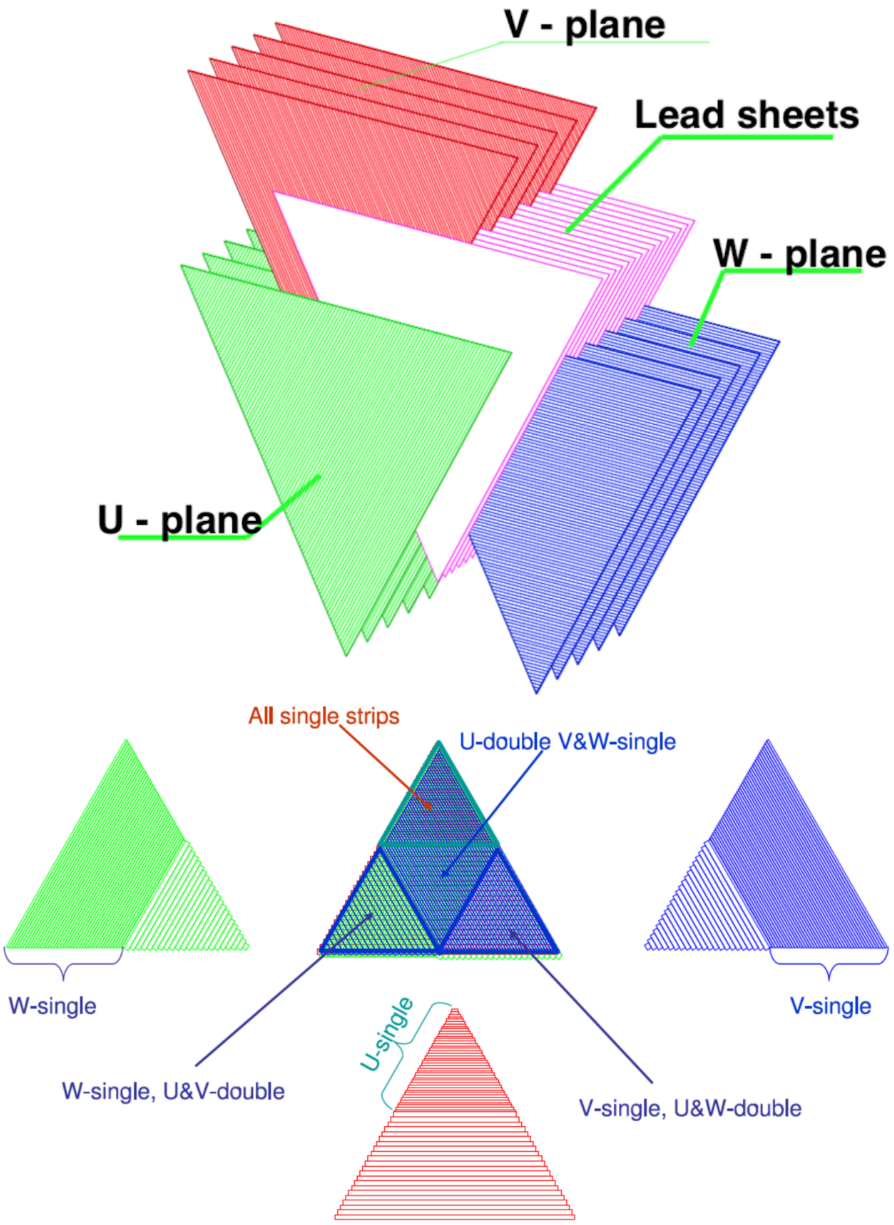


Figure 3.5: Schematic representation of the calorimeters' three views taken from [96].

The position of the cluster is then calculated based on the U/V/W peak position relative to the PMT readout end. The peak timing is corrected for the propagation delay of light to the readout using the effective velocity of light determined for each scintillator during the calibration process [96]. Calculating the time and position of each cluster is important to calculate the velocity of neutral particles as the ECAL can then be used to detect neutrals and identify them as neutrons or photons based on their velocity. The peak energies are also corrected for scintillator light attenuation before calculating the energy of the cluster. For isolated clusters this is defined as the sum of the corrected energy of each U/V/W peak. For clusters sharing a peak, the energy belonging to each cluster from a given peak is defined as the ratio of the energy from the other two peaks [103]. Measuring the energy deposition in the calorimeters is important given that photons and electrons will produce electromagnetic showers, as shown in Figure 3.6. As such, the sampling fraction defined as the ratio of the sum of the energy deposited in all three calorimeters divided by the momentum of the particle is close to a quarter for electrons and photons whilst it is typically lower for other particle types. The sampling fraction can therefore be used for particle identification, allowing to identify electrons. Figure 3.6 also contains an example of the sampling fraction of electrons. Photons are distinguished from neutrons by requiring a high velocity as measured by the ECAL. It should also be noted that the energy deposition of charged particles will vary based on the mass of the particle, as described by the Bethe-Bloch equation [104]. Furthermore, muon interactions with the calorimeter will proceed through a lepton-hadron interaction, with the muon absorbed by the constituents of the calorimeter which then emit a photon when de-exciting. Pions will interact via a hadron-hadron interaction, scattering multiple times in the calorimeters which in turn can lead to showering effects and strong tails in the energy deposition distributions. This means that the calorimeters also provides some discriminating power in terms of identifying heavier particles such as muons and pions, as will be discussed in Chapter 6.

As mentioned in the previous section, the FTOF was designed to measure the velocity of charged particles, with the identification of charged hadrons in mind. In each of the six sectors of the FD, the FTOF is comprised of three arrays of scintillator counters, referred to as panels, named panel-1a, panel-1b, and panel-2, containing 23, 62 and 5 counters respectively [95]. These are then split into left-right counterparts. As a charged particle passes through the counter it causes ionisation which subsequently generates scintillation light. The photons that are created can then be absorbed or reflected as they travel through the counter before they can impinge on the photocathode of the PMT. This interaction produces photo-electrons and the subsequent signal is amplified within the PMT. The generated pulse then inputs into the readout electronics [95]. The scintillation counters have varying length due to the geometry of the FTOF, with for example the panel-1b having counters with a length of 17 cm at the smallest scattering angles to 408 cm at the largest scattering angles [95]. These have poorer timing resolutions at larger lengths, as shown in Figure 3.7 for panel-1b, due to the increased path length taken by the scintillation light in the counter to the PMT. However, for most counter lengths the timing resolution is below the system requirement of an average time resolution of 80 ps [95]. The velocity β of a charged particle can then be estimated from the time returned by the FTOF, and the known path length to the FTOF hit position. Panel-1b is typically used for particle identification purposes due to a better average timing resolution [95]. Particle types are uniquely identified from their charge and their mass. Given that the mass m of a particle can be related to its velocity and momentum as $m = \frac{p}{\beta\gamma}$ [88], for γ the Lorentz factor, the FTOF can therefore be used to identify

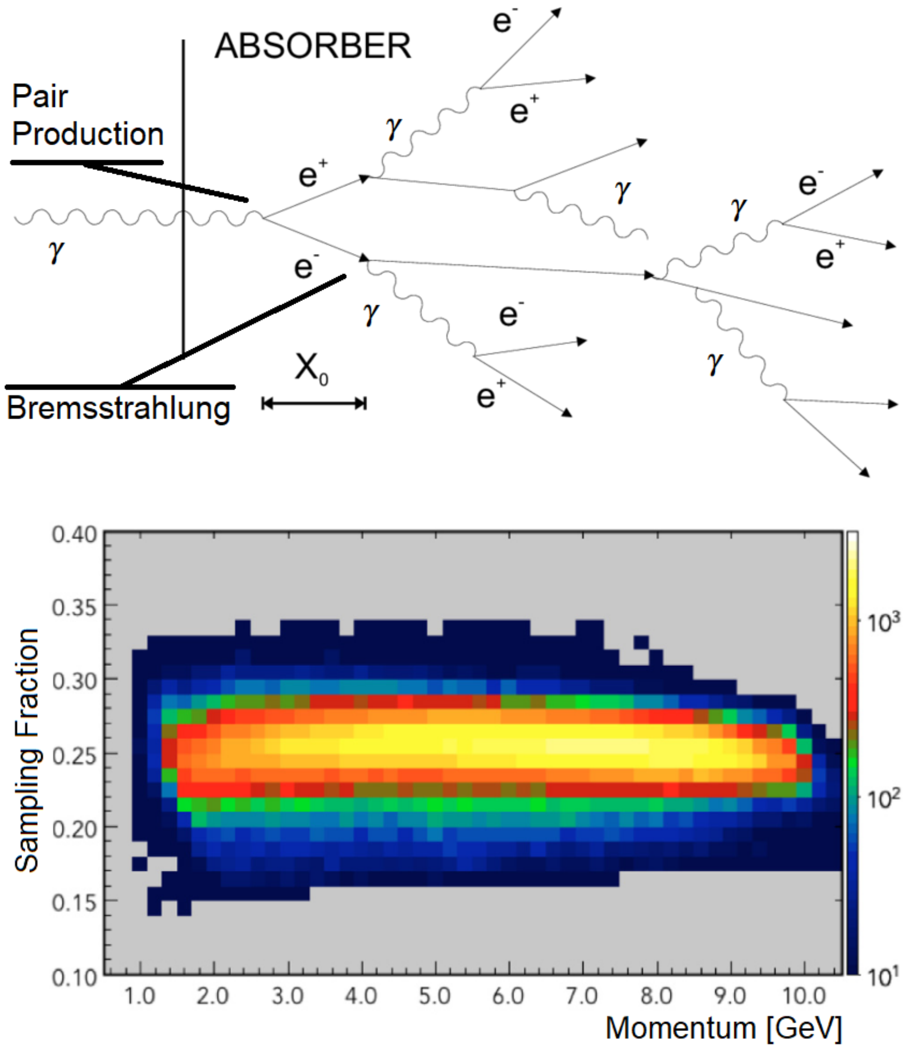


Figure 3.6: Top: Schematic representation of an electromagnetic shower in a calorimeter, taken from [105]. Bottom: The electron sampling fraction measured at CLAS12, taken from [96].

charged particles. Figure 3.7 shows the velocity as a function of momentum for three types of positively charged particles. These can easily be resolved at low momentum. The FTOF therefore provides good pion/kaon separation up to 3 GeV, and good separation between kaons, pions and protons up to 5 GeV.

The HTCC was specially designed to discriminate electrons from other charged particles. The HTCC consists of 48 mirror sections that produce Cherenkov radiation when a charged particle travels through them, with the radiated photons subsequently readout by photo-multiplier tubes (PMTs). Clusters are then made of one to four PMT signals collecting the Cherenkov light from the adjacent mirrors [92]. The number of photo-electrons produced in a cluster then becomes a discriminating variable in identifying electrons, with the requirement being at least two photo-electrons produced in the HTCC for electrons. The LTCC was built in a very similar fashion

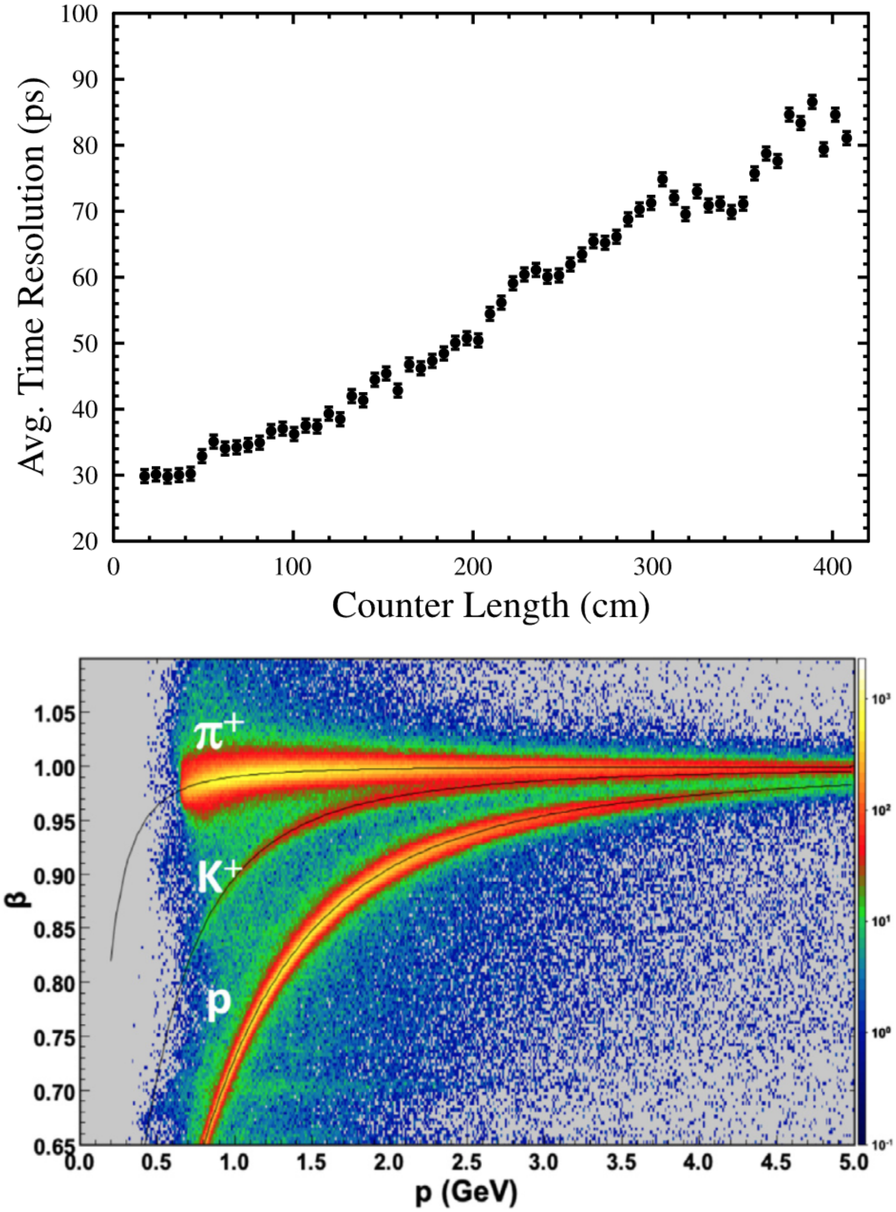


Figure 3.7: Top: The timing resolution of the FTOF panel-1b scintillation counters as a function of their length. Measurements were averaged over each of the CLAS12 sectors. Plot taken from [95]. Bottom: β versus momentum, where the velocity is measured for various positively charged particle types with the FTOF. Three distinct bands can be identified for positively charged pions (π^+) positively charged kaons (K^+) and protons (p). Plot taken from [88].

to the HTCC with 108 adjustable mirrors in each sector readout by PMTs [93]. A signal in the LTCC becomes a discriminating variable in distinguishing between pions and kaons with the lighter pions leaving a signal in the LTCC, while the heavier kaons pass through undetected. The clustering algorithm for both the HTCC and LTCC are similar, with the Cherenkov radiation produced by a particle going through the detector split into up to four PMTs for the HTCC and up to three PMTs for the LTCC. Split signals need to be combined into a single cluster to identify the signal strength in both detectors. The strongest hit in a given event is taken as the starting point for a given cluster. Adjacent hits within a given time window are then added to the cluster, and the cluster signal strength is determined as the sum of the signals in the individual PMTs. The signal time is calculated as the average between the individual signals, weighted by the corresponding number of photoelectrons, and the angular position of the cluster is taken as the simple average of the signals. The detection efficiency of the HTCC was measured at 99 % for electrons [92]. The detection efficiency of the LTCC for pions starts around 50% near the expected signal threshold at 3.5 GeV, and rises with momentum until reaching a plateau of 88% at a momentum of 5 GeV [93]. In both cases the particle identification discrimination power of the detectors relies on the fact that pions (kaons) do not produce a signal in the HTCC (LTCC). However, high momentum pions above 5 GeV will produce two or more photoelectrons in the HTCC [92], leading to a requirement for additional particle identification routines as discussed in Chapter 6. This effect is clearly shown in Figure 3.8 which shows an increase in reconstructed clusters above 5 GeV.

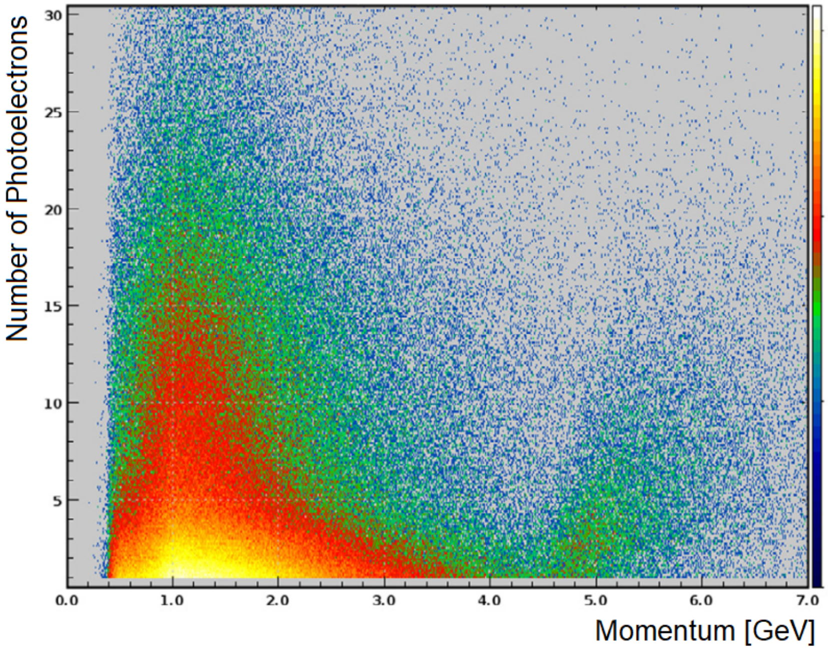


Figure 3.8: The number of photoelectrons recorded in the HTCC as a function of the momentum of the particle associated with the clusters, taken from [92]. The increase in reconstructed clusters above five GeV corresponds to high momentum pions firing the HTCC. The plot was produced for positrons and π^+.

The DC, shown in 3.9, is composed of three chambers filled with gas and containing two so called superlayers each made of six layers with 112 sense wires per layer [94], in each of the FDs six sectors. As a charged particle travels through the DC and ionises the gas in each chamber, the electrons released from the gas will be detected by the sense wires close to where they are released, mapping out the trajectory of the particle as it goes through the DC. By applying a magnetic field to the drift chamber the trajectory of the particle will curve. A positively charged particle will curve in the opposite direction to a negatively charged particle, and the amount by which it curves is related to its momentum.

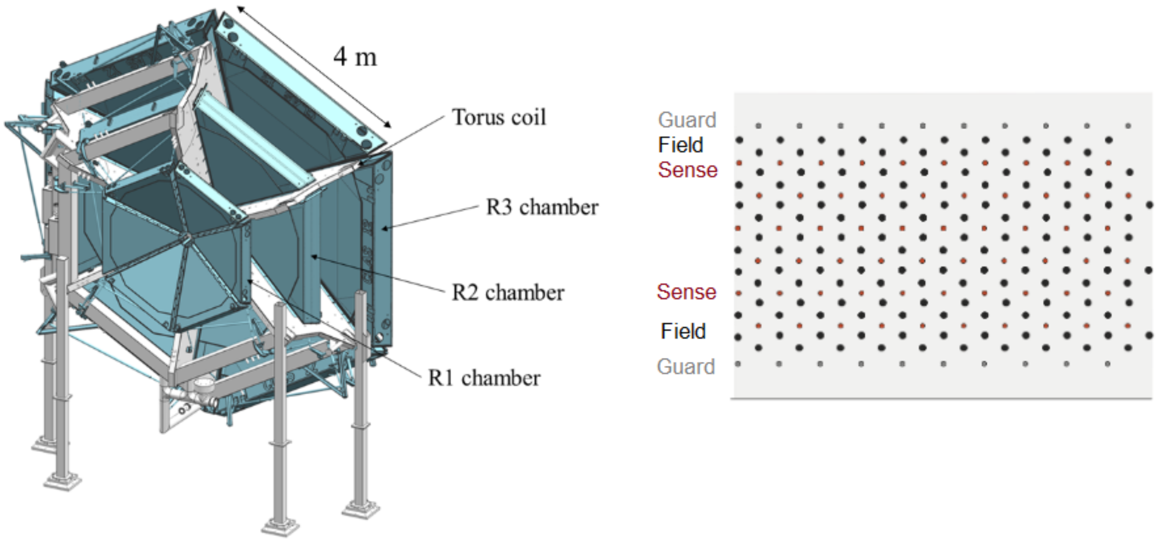


Figure 3.9: Left: Diagram showing the three chambers of the DC, each chamber containing two superlayers. The torus magnet allows to measure the charge and momentum of charged particles travelling through the DC. The forward detector's six sectors are also visible. Right: Schematic of the wire layout for one superlayer, with each superlayer containing 6 layers of sense wires (red). The view is a cut perpendicular to the wire direction. Both images taken from [94].

So called hit based tracking then reconstructs hits as a wire with a recorded signal. Groups of hits are collated into clusters identified as adjacent hits within the wire layers of a given DC superlayer [103]. Additional noise rejection algorithms are then applied to clusters with specific shapes, such as horizontal lines along a single layer, or so called curlers due to low momentum background [103]. The resulting clusters are then fit to a straight-line hypothesis over a given superlayer, and the fits with acceptable residuals are kept and the hits are identified collectively as a track segment. A Simple Noise Removal [103] algorithm is applied to the track segments, by considering the maximum tilt of a segment and requiring hits in at least three layers in every superlayer for each segment. The next step of the tracking algorithm aims to provide track candidates by first using the wire direction in a given superlayer along with the line fit to a segment in that superlayer to construct a plane. The intersection of two planes in the chambers containing two superlayers are then referred to as a cross, and a parabolic fit to a cross in each of the three chambers defines a track candidate [103]. Figure 3.10 exemplifies the track segment fit and definition of cross between two superlayers, as well as noise rejected by the pruning algorithm. A further requirement is applied to the track to have a segment in at least five superlayers in every sector. The track candidates are then fed to a Kalman Filter which fits the position and momentum parameters of a track to hits in the DC.

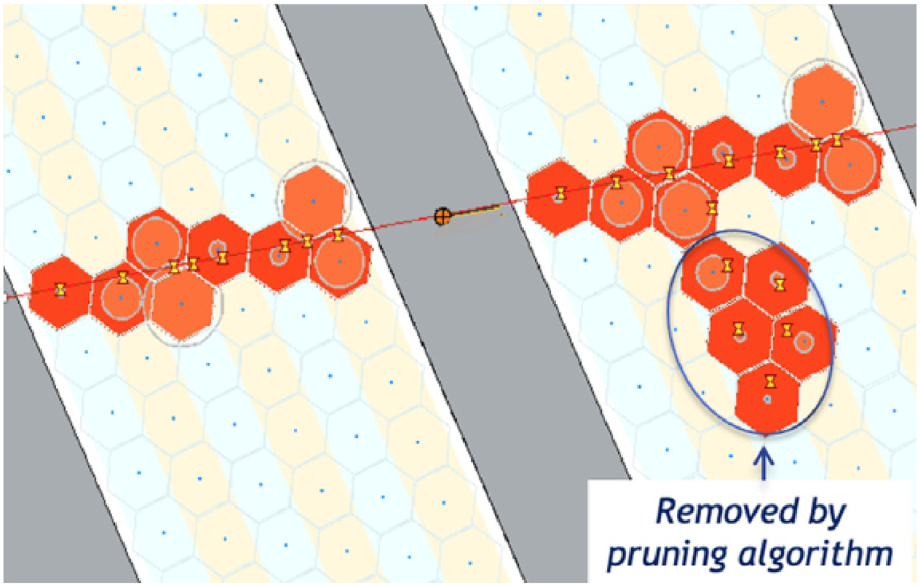


Figure 3.10: A schematic of hits in two superlayers of chamber, fitted as track segments and forming a cross. Hits along two horizontal lines on two layers were identified as noise and removed by the pruning algorithm. Plot taken from [103].

Recent efforts at CLAS12 have also lead to new machine learning based techniques for noise reduction and track candidate suggestion aiming to improve on the ones described above. In [106] a convolutional auto-encoder, as described in Chapter 5, was used to remove noise from the DC. The training sample for this denoiser was made of all hits in a given sector as the input with the clean reconstructed tracks taken as the output. In [107] it was shown that an MLP can be used to suggests new track candidates by training it on the average position of a segment in a given superlayer as the input, and the output having three nodes corresponding to a positive, negative or false track. Reference [107] also describes an autoencoder capable of identifying missing hits so as to fix the corruption in possible track candidates due to inefficiencies of drift chambers. The auto encoder was trained by taking clean reconstructed tracks and randomly removing a given hit so as to create the output and input respectively. Figure 3.11 demonstrates the gain achieved by the AI tracking algorithms at CLAS12. As shown, for a reaction $ep \rightarrow e'\pi^+\pi^-(p)$ where two pions are produced, the scattered electron e' is detected without observing the recoil proton, the tracking efficiency defined as the reconstruction of the electron and two pion tracks increases relative to the conventional algorithm when adding both the denoising and track candidate algorithms, either separately or together. This holds true when increasing beam current, which itself increases occupancy and noise in the drift chamber, typically leading to inefficiencies in tracking.

After a hit-based track has been found, the raw times of the hits are corrected based on the propagation time along the hit wire, the particle time of flight, and the cable delays [103]. A distance of closest approach (DOCA) to the sense wire is then estimated from the TDC value

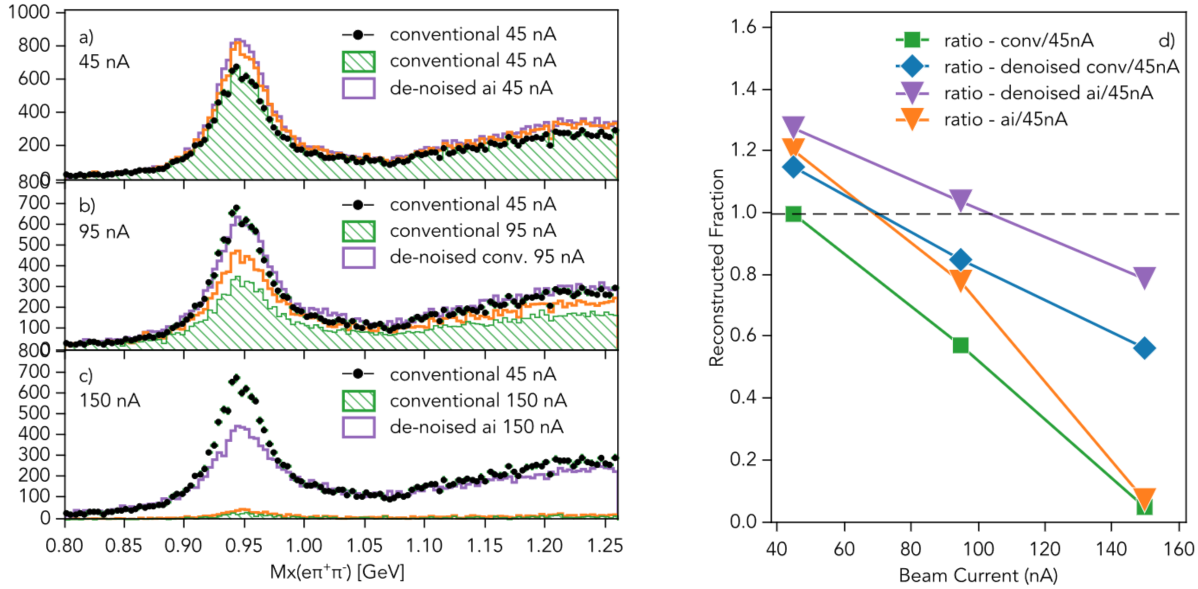


Figure 3.11: Left: The missing mass of $e'\pi^+\pi^-$ at three values of beam current, each showing a peak at the proton mass as expected. The plots compare the missing mass in data taken at 45nA and reconstructed with the conventional algorithm to data reconstructed with the conventional algorithm at a given beam current, or using denoising, or AI suggested track candidates. Right: The ratio of the tracking efficiency over the efficiency using the conventional tracking algorithms at 45 nA, for either the reconstruction using conventional tracking algorithm, the AI denoised but conventional track candidate algorithm, the AI denoised and AI suggested tracks, or just the AI suggested tracks without AI denoising. This ratio is plotted as a function of beam current. As seen, the algorithms incorporating AI all perform better than the conventional algorithm for all values of beam current. Plots taken from [108].

of the hit and timing information. This then leads to time based tracking which defines hits from the DOCA. The time based hits associated with the previous track candidates are then passed through a second iteration of the Kalman Filter [103] to determine the momentum and position of the track. The track parameters obtained from this second fit are then transformed back into the lab frame and the track is swam through both the torus and the solenoid fields to the distance of closest approach to the beamline [103]. The momentum of the particle associated with the track is then reported at this location. A final track propagation is performed to obtain the trajectory of the track as a series of points and path lengths corresponding to its intersection with all the detector surfaces [103]. This is used for subsequent matching of the track to the detector responses allowing to associate the information collated by every detector subsystem to a single reconstructed particle track, as shown in Figure 3.12.

3.3 Software & Reconstruction

As alluded to at the end of the previous section, information from each of the individual detector subsystems needs to be collated into a single track, which corresponds to a particle

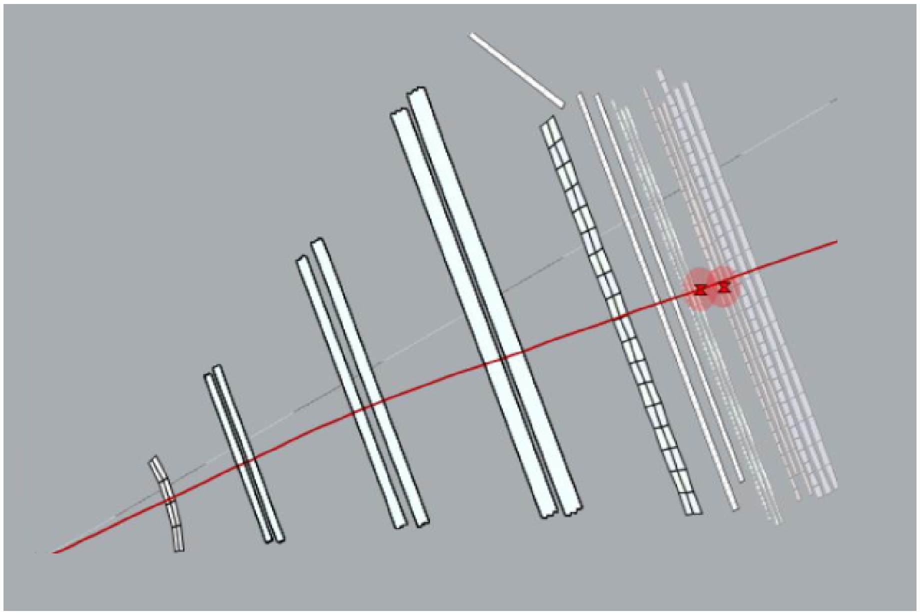


Figure 3.12: An electron track displayed in the CLAS12 Event Display (CED) [103].

going through CLAS12. To this end, the CLAS12 Event Builder (EB) collects information from the various detectors to create particles, performs a general particle identification scheme before organising the resulting information into a standardised data format. This is done on an event by event basis where an event constitutes of several detected particles within a 4ns bin corresponding to the beam bunching employed by CLAS12.

Whether or not an event is recorded is decided by the CLAS12 trigger system [109], the aim being to reduce the amount of data saved to tape by efficiently selecting events relevant to the CLAS12 experimental program whilst removing events that are not relevant. This then reduces the cost of data storage and post processing times without decreasing the amount of available and relevant data. The trigger is separated into three stages: the first stage receives information from the various detector subsystems and performs general data processing depending on the type of detector. This information is then passed onto the second stage which performs timing and geometry coincidence between the various subsystems, split into the six sectors of the FD as well as a separate coincidence requirement for the CD [109]. The third stage forms the final trigger decision, based on requirements identified for the CLAS12 experimental program. For example, given that CLAS12 operates with an electron beam, a crucial requirement is to have at least one electron detected in the FD or FT. In the FD, this is done by requiring geometrical matching between hits in the DC, HTCC and ECAL, with the same track requirements for the DC as described in the previous section, two photoelectrons produced in the the HTCC and at least 60 MeV deposited in the PCAL [109]. Appendix A.1 describes how convolutional neural networks can be employed to improve the purity of the electron trigger in the FD by removing events without an electron detected in the FD, leading to a sizable reduction in recorded data without reducing the number of events with an electron in the FD. In the FT the electron trigger requires a hit in the hodoscope that is correlated in position and time with a hit in the calorimeter [109]. To accommodate the J/ψ experimental program, an additional trigger was developed to identify events with two muons detected in the FD produced in the J/ψ decay by placing requirements on their energy deposition in the ECAL [109], as detailed in Chapter 6. The reason for doing this is that, as explained in Chapter 2, the scattered electron is not detected in the analysis of J/ψ photoproduction at CLAS12. Events where J/ψ decays to an electron positron pair will be selected by the electron trigger, whilst events where J/ψ decays to two muons will be selected by the muon trigger. After the trigger the CLAS12 Data Acquisition system reaches rates of around 500 MB/s.

Once an event passes the trigger the information from the various detectors is saved to tape, now ready for reconstruction by the EB. As described in the previous section, hits in the various detectors are collated together into clusters before being associated with a given track, for example based on the DC tracks in the FD. Clusters in the ECAL, or the CND in the CD, that cannot be associated with any tracks then form neutral particles [103]. Next a start time is associated with the event, which corresponds to the time of the interaction of the electron beam with the target at the vertex. This start time is based on the highest energy electron in the ECAL if this exists and if not with a positron in the ECAL. If neither exist the start time is taken from the forward going positive track with the highest momentum, or negative track if no positively charged track is present in the event. To facilitate analyses which require a very forward going electron in the FT, a parallel start time is also assigned based on the detection of electrons in the FT. A correction to the start time is then performed to align the event start time with the best measure of the beam-bunch arrival time at the target [103]. A time of flight is then associated to each particle, based on the start time and the time and path length to an associated cluster in a given detector for a given particle. For example, for charged particles in the FD, this time and path length would be taken from the FTOF, whilst neutral particles in the FD would use that of the ECAL. This can then be used to calculate the velocity of a particle.

The EB then performs a general particle identification scheme. First, if a charged particle produced two photoelectrons in the HTCC, has an energy deposition above 60 MeV in the ECAL and a sampling fraction within five standard deviations of the mean parameterised for electrons, then that particle is identified as an electron or positron depending on the charge measured in the DC. If a particle is not identified as an electron then it will be identified as either a pion, kaon, proton or deuteron depending on its momentum and velocity, as detailed in the previous section. Finally, for neutral particles, if the velocity of the particle is close to 1 then the particle is assumed to be a photon, otherwise it is taken to be a neutron [103].

The final stage of the EB is to collate the information from each detector subsystem, along with information from the trigger, start time and particle type identified by the EB, into a standardised bank structure. This format, called High Performance Output (*HIPO*), was developed specifically for CLAS12 [103, 110]. This was designed to provide data compression based on LZ4 and random access, whilst storing data in separate records with adjustable size, allowing events to be separated based on their contents, such as for example the number of reconstructed particles. This feature is notably used to skim the produced data output into smaller data-sets based on the requirements for one or more analyses. As such, the CLAS12 data-sets contain information for each particle in a recorded event, along with information about the event itself such as the start time or information from the trigger. Data analysis at CLAS12 then proceeds using data analysis tools like *clas12root* [111] or *chanser* [112]. The idea here is to provide tools that can read data files recorded in the *HIPO* format and simplify analyses. Appendix A.2 provides more information on both software packages and details the analysis procedures implemented towards the measurement of J/ψ photoproduction at CLAS12.

Chapter 4

Calibrations and Datasets

Careful calibrations of CLAS12 data are done to ensure an optimal reconstruction of the data as described in Chapter 3. For brevity, the calibrations will be summarised only for the Forward Detector subsystems, namely the HTCC, LTCC, DC, ECAL and FTOF.

The HTCC offline calibrations serve two main purposes: first to be able to correctly estimate the number of photoelectrons produced in the HTCC and second to align the timing of individual channels to aid in the cluster reconstruction [92]. The main issue with correctly estimating the number of photoelectrons produced in the HTCC is the high dark noise rate that hides the single photoelectron peak in a given PMT. A Light Monitoring System (LMS) was built allowing to measure the PMT response obtained with the LMS [92]. This is then used to extract the position of a single photoelectron peak when the average intensity of the LMS light is of a few photoelectrons. The results are then used for gain matching, with the calibration measurements performed for all PMTs in parallel at the same LMS settings [92]. After calibrating the single photoelectron peak position, different factors such as imperfections of the HTCC mirrors or dust deposition will lead to a variation in the signal strength from the individual channels. The signal strength in each channel is then estimated, and corresponding correction factors are used to align the signals of each individual channel to the average value between channels [92]. This then allows for a uniform response across channels, with both calibrations enabling to accurately measure the number of photoelectrons produced in the HTCC. Finally, the timing of each individual channel must be aligned to correctly reconstruct clusters in the HTCC. To this end, the time at the beam-target interaction vertex for each of the 48 channels is calculated. The time shift between the individual channels is then estimated. These time shifts then allow to align the timing of each HTCC channel [92].

For the LTCC, the main calibration aims to match the gains of each PMT. This is carried by using data taken at the very start of a given run period, so that the LTCC can be calibrated for the rest of the data taking period [93]. The calibration data is taken with a random trigger which includes LTCC events with PMT noise containing a single photoelectron peak. The ADC spectra of each PMT is fitted to find the single photoelectron peak position. The PMT high voltages are then adjusted to align the peak positions to a particular ADC value of $ADC_{SPE} = 200$. During the reconstruction, the number of photoelectrons for the digitized ADC value is then calculated as $\frac{ADC}{ADC_{SPE}}$ [93].

As mentioned in Chapter 3 the second step of the DC reconstruction is so called time based tracking based on the distance of closest approach (DOCA) estimated from the wire TDC. The TDC to DOCA conversion needs to be carefully calibrated to ensure good reconstruction of the tracks in the DC [94]. A 4th order polynomial is used to model the distance to time parametrization. Physical constraints are then placed on the drift velocity of the ions inside the DC gas chambers, enabling to relate the parameters of the distance to time function to physically meaningful parameters [94]. The dependence of the distance to time function on the local angle of a hit and on the magnetic field strength then allow to correct some of these parameters based on the local angle and magnetic field strength. Finally by using the hit based track to calculate a track distance to closest approach (TRKDOCA), a fit to the dependence of the time as a function of TRKDOCA determines the values of the parameters of the distance to time function. This function is then numerically inverted to estimate the DOCA from the TDC value of the sense wires [94].

The ECAL calibrations aim to calibrate the energy deposition and timing measurements of the ECAL [96]. The energy parametrization require a parameterization of the light attenuation as a function of the distance from the cluster to the readout end of the scintillator strip [96]. Minimum-ionizing particles (MIPs) such as cosmic muons are used due to their relatively simple energy deposition profile allowing to determine the light attenuation parametrisation for each PMT during engineering runs with cosmics before the start of data taking. The PMT HV can then be adjusted to produce a uniform overall response [96]. During the cosmic runs, a trigger ensures that only clusters that correspond to the overlap of single strips in each of the ECAL's three views have been selected. These clusters are called pixels, and are the simplest possible cluster corresponding to the smallest positional resolution of the ECAL, allowing to minimise the spread of the MIP energy deposition [96]. The energy loss distribution for each pixel for a single PMT can then be measured on a per pixel basis, allowing to measure the attenuation as a function of distance to the readout end of the strip. This is then fitted to extract the desired parametrisation [96]. Optimal timing calibration in the ECAL is necessary both for neutral particle identification and background rejection, as neutral particle identification relies on measuring the velocity of the particles using the ECAL. By using the path length to the ECAL and the start time of the event measured in other parts of the reconstruction, an expected readout time can be established for charged particles like electrons and pions [96]. A χ^2 minimisation is then performed to minimise the difference between the measured and expected time, allowing to parametrise calibration constants such as the scintillator effective velocity or the amplitude dependence of the PMT pulse time [96].

Careful calibrations of the FTOF are required to achieve high resolution in charged hadron identification. As the FTOF calibrations rely on accurate path length measurements to the FTOF, the FTOF is calibrated after the drift chambers calibrations. First a time offset between the left and right PMT counterparts accounts for the difference in time measured by the left or right PMTs [95]. Next an ADC calibration determines the ADC value to energy deposition calibration factor for each counter using MIP events [95]. The attenuation length calibration of the scintillation counters quantifies the light absorption length in each counter. This is determined by measuring the ADC as a function of hit coordinate along the bar [95]. Effective velocity calibration allows to determine the effective speed of light propagation along the counter, to then correctly measure the time of a hit in the counter [95]. Time-walk amplitude calibrations then compare the measured hit time to the measured ADC to determine the time-walk correction, also taking into account the position of the hit [95]. Counter to counter time offsets are then calibrated in a similar manner to the HTCC to provide consistent timing information between counters [95]. Finally, the TDC calibration allows to establish the channel to time calibration. This is mainly done as a cross check as a single conversion factor is employed for all TDC channels [95].

Once a dataset is calibrated, it is then ready to be used for analysis. Two different datasets taken by CLAS12 are used for the analysis of J/ψ photoproduction at CLAS12. The first, so called RG-A fall2018 was taken in the fall of 2018 with a liquid hydrogen (LH_2) target. As the nucleus of a hydrogen atom only contains protons, the protons in a hydrogen target are usually considered as free protons, allowing to measure J/ψ photoproduction on the (free) proton. 75% of the fall2018 dataset was taken with a 45 nA beam current, with 8% at 50 nA and 17% at 55 nA. As explained in Chapter 3, the beam current affects the reconstruction efficiency as it will increase the occupancy in the drift chambers and calorimeters and lead to additional noise which hampers the reconstruction algorithms. The fall2018 dataset was taken with a 10.556 GeV beam, with a total accumulated charge of 37.06 mC.

The second dataset, so called RG-B spring2019 was taken in the spring of 2019 with a liquid deuteron (LD_2) target. In this case the nucleus contains bound protons and neutrons, allowing to measure J/ψ photoproduction on the bound proton or neutron. Due to an issue with the accelerator, the spring2019 dataset was taken at two different beam energies. Roughly 40% of data was taken at 10.5986 GeV, with roughly half taken with a 35 nA beam current and the other half taken at 50 nA. The other 60% of the spring2019 dataset was taken at 10.1998 GeV with a 50 nA beam current. The accumulated charge at 10.5986 GeV is of 27.01 mC, with the accumulated charge at 10.1998 GeV being 39.39 mC.

The accumulated charge is used to calculate the number of photons produced during data taking, which forms a part of the luminosity calculation as detailed in Chapter 8. Another part of this calculation is the number of protons or neutrons in the target. This is calculated as:

$$\rho_T = N_N = 2N_{molecules} = 2N_A \frac{\rho}{M} \tag{4.1}$$

for ρ_T the target density equals to the number of nucleons N_N, which is twice the number of molecules $N_{molecules}$ of LH_2 or LD_2, N_A Avogadro's number, ρ the physical density and M the molar mass. As an example, for deuteron, the physical density is $\rho = 0.163$ g cm^{-3} with the molar mass $M = 4.028$ g mol^{-1}. The number of protons or neutrons in the LD_2 target is therefore $\rho_T = 4.87 \cdot 10^{22}$ nucleons per cm^{-3}. For the LH_2 target this is $\rho_T = 4.14 \cdot 10^{22}$ protons per cm^{-3}. The length l_T of both the LH_2 and LD_2 targets is $l_T = 5$ cm.

A pre-skim, so called train, of both datasets is made to reduce the amount of data processed during analysis. For the channels with a proton, electron and positron, the requirement is that both electron and positron are present in the forward detector (FD), and to have one proton in either the forward or central detectors. Any number of additional charged or uncharged particles are allowed. The definition of electron, positron and proton is given in Chapter 6. For channels with a proton and di-muon pair, the requirement for the proton is the same, and for the muons the requirement is that there are two muon candidates in the forward detector selected with loose cuts on the energy deposition as explained in Chapter 6. Note that these energy deposition cuts were originally too restrictive and were loosened for the RG-A fall2018 dataset but not for the RG-B spring2019 dataset. As such, for the moment, the analysis of J/ψ photoproduction with J/ψ decaying to a di-muon pair is limited to the RG-A fall2018 dataset. For the channels with a neutron, the pre-skim requirement is simply to have one electron in the forward detector and any number of additional charged or uncharged particles.

A final consideration with regards to the datasets is that a new version of the reconstruction software will soon be deployed for both datasets in a so called second pass over the datasets. This will include the AI improvements to the drift chamber reconstruction, as mentioned in Chapter 3, along with other changes to the reconstruction, and is expected to give a roughly 40% increase in statistics with three charged particles in the FD, and roughly 30% for only two charged particles. As such this will lead to a significant increase in the statistics available for the analysis of J/ψ photoproduction at CLAS12, which is limited by the low J/ψ yield. On the other hand, this will also mean that all simulated data will need to be regenerated with the new reconstruction software, and some analysis procedures will need to be adapted. The pre-skim energy deposition requirements for muon candidates will also be loosened for the RG-B spring2019 dataset. The results presented in this thesis are therefore preliminary, indicative that interesting results can be obtained from J/ψ photoproduction at CLAS12, and will be refined once this new pass is produced.

The simulated data is produced with a Geant4 [113] based simulation software called *GEMC* [114] which is specific to CLAS12. This allows to recreate the detection of particles with CLAS12 and the same reconstruction algorithm is applied to the simulated data. Background merging allows to emulate data taken at given beam currents. The background is typically taken from random trigger data which was initially identified to emulate physics and electronic backgrounds in the various detectors when simulating the CLAS12 detector [114]. This then allows to simulate the reconstruction efficiency which is decreased as the beam current increases due to increased backgrounds. *GEMC* will simulate the response of the detector to the events which are passed to *GEMC*. The inputs to *GEMC* are often called true or generated, as these contain the true four-momentum of the final state particles and not the four-momentum of the final state particles as reconstructed by the simulation. These reconstructed four-momenta are modified with respect to their true values to account for the resolution of the detector. The true events are generated with *elSpectro* [115], an all purpose event generator which allows to correctly generate photo and electro production of mesons and baryons at experiments such as CLAS12 but also GlueX or others. Figure 4.1 shows a comparison of t and the quasi-real photon energy for the RG-B spring2019 data and simulated data in the $en_{bound} \rightarrow (e')e^+e^-n$ channel. These show good agreement, demonstrating that the simulation produces a realistic representation of CLAS12 data. elSpectro will simulate the Fermi motion of the bound constituents of the nucleon and can either be set up to use the JPAC predictions for the J/ψ photoproduction cross section, or without placing constraints on the kinematics of the J/ψ decay. The later option was used here for comparison to the data taken with CLAS12. The $en_{bound} \rightarrow (e')e^+e^-n$ channel was chosen here for comparison as this requires the most corrections to the four-momentum of final state particles and would be the most susceptible to differences between simulation and CLAS12. The simulations are then produced at different beam energies and with background merging corresponding to different beam currents, with the same proportion of each as to that in CLAS12 data. This amounted to roughly 5 million events in total for each of RG-A fall2018, RG-B spring2019 at 10.2 GeV and at 10.6 GeV. Additional simulated data will be produced once the new pass has been deployed.

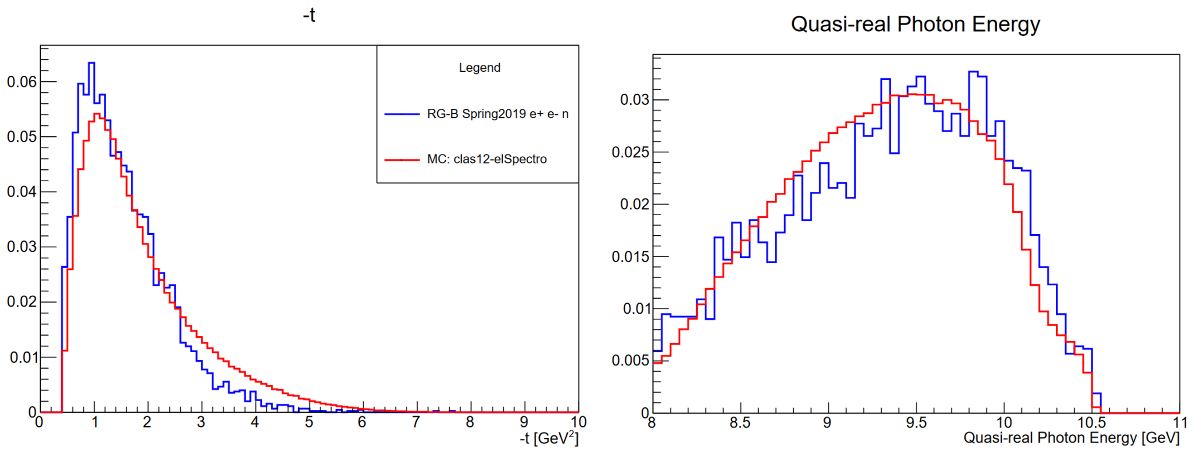


Figure 4.1: t (Left) and the quasi-real photon energy (Right) for RG-B spring2019 data (blue) and simulated data (red) in the $en_{bound} \to (e')e^+e^-n$ channel.

Chapter 5

A Users Guide to Machine Learning

Machine learning has become more and more prevalent in high energy physics since the start of the century, with an explosion of applications in topics like particle identification or reconstruction in the 2010s [116]. However, smaller collaborations like those based at JLab are still catching up with larger collaborations like those based at the Large Hadron Collider (LHC) that typically have the required manpower and funding to investigate and deploy cutting edge technologies within shorter time frames. The analysis of near-threshold J/ψ photoproduction at CLAS12 proved to be an ideal testing ground for machine learning based analysis techniques aiming to improve particle and reaction identification, due to a decreasing cross section in the near-threshold region which led to small event samples compared to sizeable backgrounds in some of the channels that were investigated for this thesis. Similarly, Appendices A.1 and A.3 describe projects that aimed to investigate the use of machine learning to create an AI electron trigger at CLAS12 and machine learning based fast simulation routines. A good understanding of machine learning will be absolutely necessary to follow parts of the discussion in later sections of this thesis. There are however many detailed and engaging resources available in multiple formats [117–119] that can give a much more thorough and digestible discussion of machine learning than could be achieved within a single chapter of a thesis in hadronic physics. As such, instead of aiming to re-invent the wheel, this chapter will aim to provide the reader with a good understanding of how to use machine learning by describing the general layout and components of a machine learning based project and the various algorithms that will come into play throughout this thesis. This chapter will link to an online tutorial on removing noise from a toy drift chamber so as to exemplify the points made during the discussion. The tutorial was made to give the opportunity for a hands on exercise on using machine learning during an informal seminar on the use of AI at CLAS12.

5.1 The Machine Learning Pipeline

Machine learning can be loosely described as a set of tools or algorithms that leverage data to create predictive models without having to explicitly code in these models [117]. For example, a *hello world* problem in machine learning is teaching an algorithm to recognise the handwritten digits that compose the Modified National Institute of Standards and Technology (MNIST) dataset. In this case an algorithm is given examples of handwritten digits before creating a model that predicts what digit is passed to the algorithm, without requiring the user to code in the model allowing the algorithm to recognise digits.

Machine learning algorithms can be categorised based on how these learn and what their objectives are. All the machine learning algorithms used in this thesis are examples of supervised learning. In this case, the dataset passed to the algorithm during training contains a feature matrix $\mathbf{x}$ and an associated label matrix $\mathbf{y}$ with an equal number of rows [117]. The training process then teaches the algorithm to predict each of the labels contained in a row of $\mathbf{y}$, for each corresponding row of the feature matrix $\mathbf{x}$ [117]. The feature matrix contains variables, known as features, that allow the algorithm to make this prediction. Once the algorithm is fully trained, it is then deployed on unlabelled data, where $\mathbf{y}$ is unknown. The opposite of supervised learning, unsupervised learning, uses only the feature matrix $\mathbf{x}$ to train the algorithm which then outputs an unspecified label matrix [117].

As only supervised learning algorithms are used in this thesis, from now on the term *machine learning* can be assumed to refer to supervised learning. In addition, the terms AI and machine learning are used interchangeably. Machine learning is a subset of AI but both terms can here be understood to refer to the same algorithms. As mentioned, the goal of a supervised learning algorithm is to produce a label y subset of the matrix $\mathbf{y}$ for each element of the feature matrix $\mathbf{x}$. In practice, the terms *classification* and *regression* are used to distinguish between the cases where the set of values in $\mathbf{y}$ are discrete or continuous [117]. Classification can be thought of as trying to label each row of $\mathbf{x}$ into certain classes. The MNIST example given above is an example of a classification problem, where each handwritten digit is labeled into one of ten classes corresponding to the digits ranging from zero to nine. Each row of $\mathbf{y}$ will then contain ten numbers which are the labels for each class. Each row of $\mathbf{x}$ contains a handwritten digit and each row of $\mathbf{y}$ contains a label set to one corresponding to that digit. For example, when a row in $\mathbf{x}$ contains the digit 0, the associated row in $\mathbf{y}$ would be $[1,0,0,0,0,0,0,0,0,0]$. Appendix A.1 also gives an example of a classification task where events detected by CLAS12 are labelled as having an electron or not, which then allows to trigger on having or not an electron in an event. This can be thought of as a particle identification task, where hits in the detector are labeled as belonging to an electron or not. Regression problems aim to predict a number associated with each row of the feature vector. For example, Appendix A.3 details how machine learning can be used to model the resolution of a given detector. In this case, the algorithm predicts the deviation introduced by the detector from the true four-momentum of a particle. Another example of a regression task is image generation which aims to create new images from a given input.

As is often the case when learning about new programming tools, the best way to get to grips with machine learning is to try it out. A short online tutorial will accompany the discussion in this section, where the aim of the tutorial is to use machine learning to remove noise from a toy drift chamber meant to resemble that of CLAS12. Chapter 3 contains more detail on how a drift chamber works, but for the purpose of this discussion it is sufficient to define this as a set of wires arranged in consecutive superlayers that allow to track a charged particle as it goes through the drift chamber. Denoising a drift chamber is a regression task where the algorithm will aim to generate a noiseless image of a track in the drift chamber from a noisy image that contains the track and noise. The tutorial is contained in a Jupyter Notebook which can be found here [120].

The machine learning pipeline generally refers to the set of steps needed to train and deploy a machine learning algorithm. The first step towards using a supervised machine learning algorithm is therefore to create a dataset on which to train the algorithm. This dataset is typically separated into training and testing sets. These ensure that the algorithm is not overfitting to the training set [117], that is to say that it is not creating a model so specific that it only applies to the dataset used to train the algorithm. This can be checked by applying the algorithm to a testing dataset: if the algorithm performs poorly on the testing set but performs well on the training set then it is overfitting, the expectation being that the model is generalisable to unseen data. Typically the same data is used to create both the training and testing sets, although this can sometimes lead to errors in the dataset going unnoticed.

Creating the training dataset might be the most important step in the pipeline as ultimately, the algorithm can only learn based on what it is given to train on. It is also relatively easy to introduce some kind of bias into a training set. A famous example [121] is that of an algorithm trained to distinguish between images of huskies and images of wolves. At first glance, the algorithm seems to perform well, even when applying it to a testing set which was created in a similar fashion to the training set. However, when deployed to new images that were not used in the training or testing set the algorithm did not perform as well. The reason? In both the training and testing sets, the images of wolves always contained snow in the background, and the algorithm was actually trained to recognise snow [121]. A common example in hadronic physics is seen when trying to train an algorithm to distinguish between two different particle types, which is commonly known as particle identification (PID). A bias is introduced when creating the dataset due to one of the particle types being measured over a different momentum range to that of the other particle type. As the response of a lot of detector subsystems correlates to momentum, the algorithm then learns to distinguish momentum ranges even when not explicitly trained on momentum, and associates a given particle type to a given range. As such the algorithm is biased towards momentum, and when deploying it, it will incorrectly classify particle types based on their momentum.

A common way of creating training datasets in hadronic physics is to use simulated data. This makes it easier to remove any biases from the training data, as it is easier to know exactly what goes into the dataset, and it also makes it easier to have representative data. For example, for PID, the training dataset must only contain examples of the particle types that the algorithm must identify, otherwise it is unrepresentative of the problem at hand. To create such a dataset from data taken with a real detector, identifying the particles that will go into the training data must be done with absolute certainty. Whilst this is feasible, it will most often rely on using a specific reaction to identify this particle type. Assuming this can be done for all particle types that are to be identified by the algorithm (over the same momentum ranges), it might not be the case that the kinematics of these reactions will correspond to the kinematics of the events in the preferred use case. As such the training data would be unrepresentative of the task at hand, which might lead to a decreased performance when deploying the algorithm. The drawback from using simulated data is that this may not always be fully representative of data taken with the actual detector, and care must be taken in using features that are well simulated. Typically the easiest way to check this is to deploy the algorithm trained on simulated data onto data taken with the real detector that the simulation aims to recreate. If the algorithm performs well on both then the simulation must be representative of the real detector.

As stated the aim of the accompanying tutorial is to remove noise from a toy drift chamber. The training data is then created artificially for this toy drift chamber: the first few cells of code of the tutorial are aimed at generating data to mimic a drift chamber with an arbitrary number of superlayers, wires and particle tracks per event. Each event is then described by a an array of size $S \times W$ where W is the number of wires in each superlayer and S is the number of superlayers in the toy drift chamber. This array is filled with one if a wire contains a hit and zero if it does not, which then draws out the track left by a charged particle in the toy drift chamber. The leftmost image of Figure 5.1 shows an example of an event with a clearly visible track going through a drift chamber with 20 wires in each of the 20 superlayers. These parameters, along with the maximum number of tracks per event, can then be varied as seen fit. As a comparison, the CLAS12 forward detector drift chambers are composed of six superlayers, with 112 wires per superlayer.

To mimic noise in the drift chamber, random hits can then be added to the previous arrays for each event, with a user defined maximal number of noisy entries in the array. Similarly,

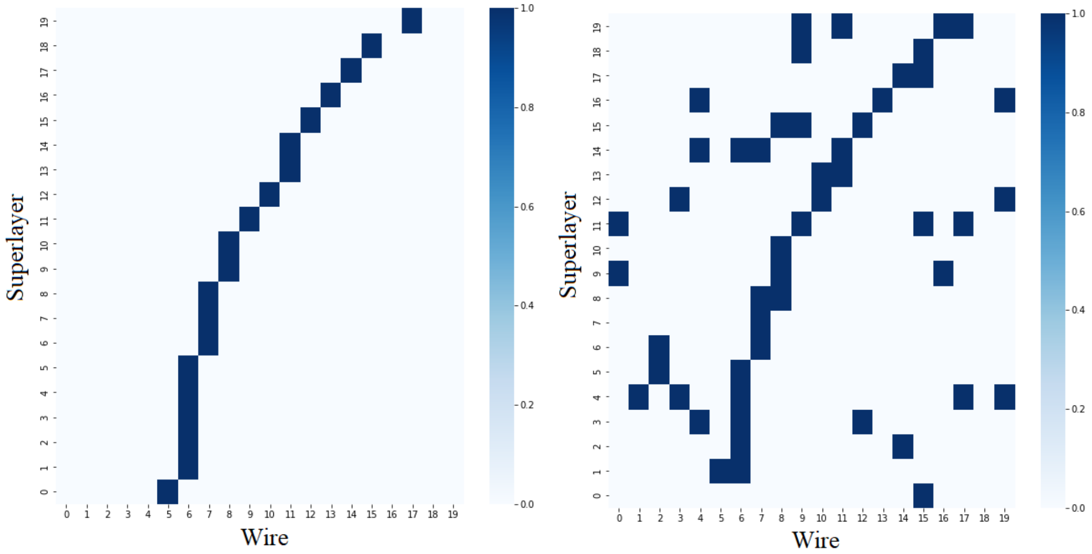


Figure 5.1: Left: An example of a track in a toy drift chamber with 20 superlayers and 20 wires. Right: The same track with noise added and a hit removed from the first superlayer. Note that the hit in the first superlayer was removed from the track before adding noise to a different wire in the same superlayer.

to mimic inefficiencies, an entire row or superlayer can be removed from the array (before or after adding noise). The rightmost image of Figure 5.1 shows the same event as before after removing a superlayer then adding noise to the event. Multiple events as the one shown in Figure 5.1 are then split into a training and testing set, which will allow to check for over-fitting of the model to the training data. The aim of this project is therefore to go from the right to the left of Figure 5.1, essentially de-noising the data from the toy drift chamber. The feature matrix **x** therefore contains M rows of noisy images, whilst the label matrix **y** contains M rows of noiseless images. Not much care needs to go into creating unbiased or unrepresentative data in this case, as the algorithm will not actually be deployed. However, if it were, then unrepresentative data could be created by giving the wrong shape to the toy detector relative to the algorithm's projected use case. Bias could be introduced by adding tracking inefficiencies before adding noise, which could lead to removing as noise all hits in a superlayer that seems to be absent in an obvious track.

An important aspect of creating a training dataset is the choice of variables, so called features, that go into the dataset. In the case of the de-noising tutorial this is rather straightforward: the input features are simply a zero or one contained in each element of the input array depending on if there is a hit in the wire at a given superlayer. This, however, can be more complicated when for example trying to do particle identification. In the case of the CLAS12 detector, as described in Chapter 3, there are many subsystems that could contribute meaningful information towards identifying the final state particles detected by CLAS12. Including too many features may hamper the training progress as this can lead to dimensionality issues whereby the algorithm will be slower to train and to make predictions as it needs to treat a larger amount of information. Superfluous variables that do not contain any relevant information for the task at hand can also lead to correlations in the training data that can introduce some bias in the algorithm. In practice relevant features can often be selected using appropriate domain knowledge. For example, some CLAS12 subsystems were built to identify specific particle types and would not be useful to identify other particle types. Feature selection techniques can allow to rank features with the aim in mind to remove the lowest ranked features. Dimensionality reduction techniques can allow to remove highly correlated features that contain redundant information, or to construct new features from highly correlated ones that contain all the relevant information.

The next step in the pipeline is to choose and train a learning algorithm. These come in several different types, such as decision trees and ensembles of these or neural networks. Certain algorithms are better suited to certain tasks, for example a type of neural network called autoencoders are known to be effective for image generation tasks. The next section will detail how the different kinds of learning algorithms are structured and how they learn to predict the label matrix $\mathbf{y}$ for a given feature $\mathbf{x}$. Suffice to say that the training process optimises certain variable components of the algorithm so as to optimise their performance. This is known as the learning process as the algorithm optimising itself is the reason why machine learning is called as it is. The performance of an algorithm is evaluated using a cost or loss function that typically estimates the deviation between the algorithm's prediction $\mathbf{y}'$ and the true labels $\mathbf{y}$. For example, for a label matrix $\mathbf{y}$ of size $N \times 1$ which has one label to predict for each of N predictions, and an associated prediction matrix $\mathbf{y}'$ the mean squared error (MSE) is defined as:

$$MSE = \frac{1}{N}\Sigma_i^N(\mathbf{y}_i - \mathbf{y}'_i)^2 \tag{5.1}$$

The example in the tutorial is an autoencoder using the MSE as a cost function. The next section will detail how to set up an autoencoder and the exact details of the one used in the tutorial. During the learning process, the algorithm will try to predict the noiseless images of $\mathbf{y}$ from the noisy images $\mathbf{x}$. In the case of an image, the MSE is summed over each pixel in the image. The MSE is then calculated for each image of $\mathbf{y}$ over several passes called epochs. The algorithm is then optimised so as to minimise the MSE for each image, with the optimisation happening at different intervals depending on the algorithm but such that each element of $\mathbf{x}$ has at least one opportunity to contribute to the optimisation. Figure 5.2 shows the MSE loss evaluated at each epoch on the training and testing sets of the tutorial dataset. The loss decreases before reaching a plateau, indicating that the algorithm has converged and finished training for both the training and testing datasets. However the difference between the loss evaluated on the testing and training is indicative that the algorithm is slightly overfitting. This is not a major concern as the loss on the testing dataset does not further increase with additional epochs which would denote that the performance of the algorithm is getting worse as the algorithm trains for additional epochs.

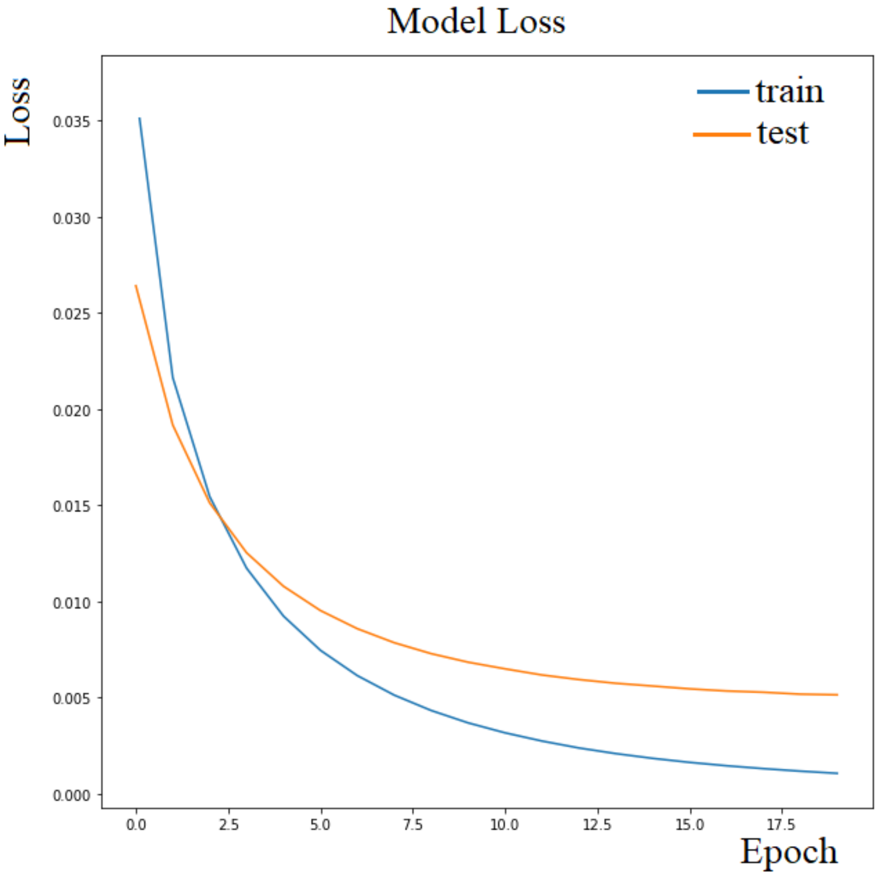


Figure 5.2: The *MSE* loss evaluated at each epoch on the training and testing datasets. The training and testing losses can be used to check for overfitting. As the testing loss is here slightly worse than the training loss, this is a sign that the model is overfitting to the training data as it performs less well on unseen testing data. However, this is not too concerning here as the testing loss does not start to increase again which would demonstrate the model performance becoming worse on unseen data as it trains for longer.

When looking at Figure 5.2 it might be hard to know whether the loss achieved after 20 epochs is good or not, which would indicate that the algorithm performs well. For image generation tasks such as denoising, an easy way to assess this is simply to look at examples of the denoised tracks. The top row of Figure 5.3 contains the same example taken from the testing set as in Figure 5.1. The bottom row of Figure 5.3 shows the output of the autoencoder when applied to the rightmost image of the top row, once the autoencoder has been trained for 20 epochs. The bottom left image is the raw output of the autoencoder, where values in each wire of the toy drift chamber range from zero to one. The raw output of a learning algorithm is called the response of the algorithm. For hits in the original track the response is close to one, whereas in other wires the response is close or equals to zero. For some wires previously containing noise close to the original track, the response is higher, closer to 0.1 or 0.2. A threshold of 0.4 is applied to the raw output, converting each wire of the toy drift chamber with a response inferior or equals to 0.4 to zero, and each wire with a response above 0.4 to one. The resulting image is shown in the bottom right image of Figure 5.3. As shown this denoised track is identical to the original track in the top left image of Figure 5.3, with the missing hit in superlayer 0 recovered and all noise removed.

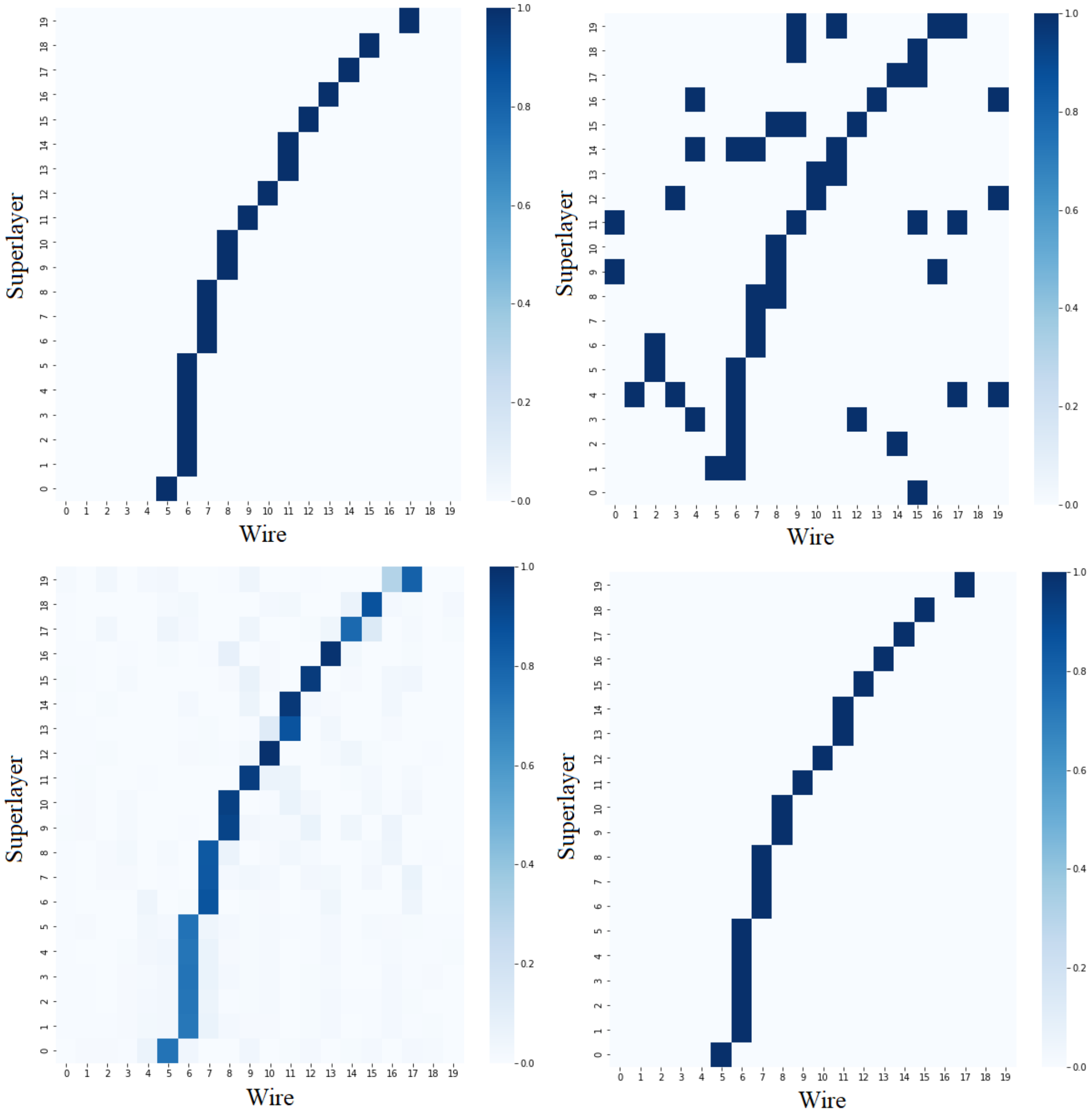


Figure 5.3: Top Row: An example taken from the original testing dataset. Left: An example of a track in a toy drift chamber with 20 superlayers and 20 wires. Right: The same track with noise added and hits in a superlayer removed.
Bottom Row: The predicted denoised image from the same example. Left: The raw output of the autoencoder. Right: The output once a threshold of 0.4 has been applied. Each wire of the toy drift chamber with a response inferior or equals to this threshold is set to zero, and each wire with a response above the threshold is set to one.

Visual checks such as the above can be helpful in qualitatively estimating the performance of a learning algorithm, but are not necessarily satisfying as they only allow to check an image one at a time without giving a quantitative assessment of the performance. The tutorial at this point is open ended; one could think of several different way to assess the performance of the autoencoder. For example, by recording the amount of noisy hits removed one can estimate how good the algorithm is at removing noise. By instead recording the amount of good hits coming from a track one can assess how good the algorithm is at preserving tracks. This might be a more important consideration, if for example the denoising autoencoder would be deployed in an experiment which necessitates a high detection efficiency and cannot afford to loose good tracks. Finally one might consider looking at fake hits introduced by the denoiser close to the original track. These could change the shape of the track, which when fitted to measure the track's momentum, would lead to a wrong value of momentum. For example in Figure 5.3, if the hit in superlayer 19 was misplaced to the left or right by one or two wires this would change the measured momentum. Once one or several metrics have been established, these could also be used to refine the threshold used to create the bottom right plot of Figure 5.3.

This notion of defining metrics to assess where to place a threshold on the response takes on a crucial role when considering classification problems. As the response of a learning algorithm for each class will not be exactly zero or one but will lie somewhere in between, a decision must be made as to where to place this threshold on the response. To exemplify the importance of this, let's consider briefly the AI trigger project described in appendix A.1 and [122]. The aim of the trigger is to select events with an electron in a certain region of the CLAS12 detector and filter out those events that do not contain an electron. A confusion matrix can then be created as in Table 5.1. A True Positive (TP) is then defined as an event with an electron that has been correctly selected by the trigger. A False Positive (FP) is then defined as an event which does not contain an electron but was incorrectly selected by the trigger. A False Negative (FN) is an event with an electron which was incorrectly rejected by the trigger. A True Negative (TN) is an event without an electron that was correctly rejected by the trigger [122].

Confusion Matrix	Electron in Event	No Electron in Event
Selected by Trigger	True Positive (TP)	False Positive (FP)
Rejected by Trigger	False Negative (FN)	True Negative (TN)

Table 5.1: The electron trigger confusion matrix. Table taken from [122].

In a binary classification problem there are two classes, positive and negative, with the response for each class equals to one minus the response in the other class. As such one can only consider the response in one class as the response in the second class would only provide redundant information. Preference is usually given to the response in the positive class for which a perfect algorithm would assign a response of 1 to an event with an electron, and a response of 0 to an event without an electron. As no learning algorithm is perfect, the response will range from 0 to 1. To determine whether or not an event should be selected by the trigger, a threshold is placed on the response of the learning algorithm forming the trigger. An event with a response above the threshold is kept whilst an event with a response below the threshold is rejected [122]. Metrics such as the purity (P), efficiency (E) and accuracy (A) can then be calculated from the entries in the confusion matrix:

$$E = \frac{TP}{TP+FN} \tag{5.2}$$

$$P = \frac{TP}{TP+FP} \tag{5.3}$$

$$A = \frac{TP+TN}{TP+FP+FN+TN} \tag{5.4}$$

A high efficiency is achieved by selecting most events with at least one electron. Misclassifying events without an electron in one of the six sectors, for example mistaking other particle types for electrons, decreases the purity of the trigger. Typically a balancing game can be played between both quantities as one decreases as the other increases, when, for example, easing or tightening the requirements on electron identification. This is reflected by the overall accuracy which is highest when maximising the product of the purity and efficiency. Figure 5.4 shows in the top histogram the response of the learning algorithm for events in the testing set that contained or not an electron, respectively in blue and red. In the bottom graph of Figure 5.4 the metrics defined above are plotted at different thresholds on the response.

The second plot of Figure 5.4 can then inform the decision making in choosing where to place a threshold on the response. In the case of a trigger, the preference is generally given to maximising the efficiency by placing a threshold at low values of the response, as ideally the trigger should preserve all good data, at the expense of introducing some background. However, in a scenario where good data was cheap and easy to produce, one might choose to maximise the purity by placing a cut at a high value of the response so as to remove most of the background, which will make analysing the data easier. Maximising the accuracy would then achieve a sweet spot in maximising both the purity and efficiency. More detail will be given on the AI trigger

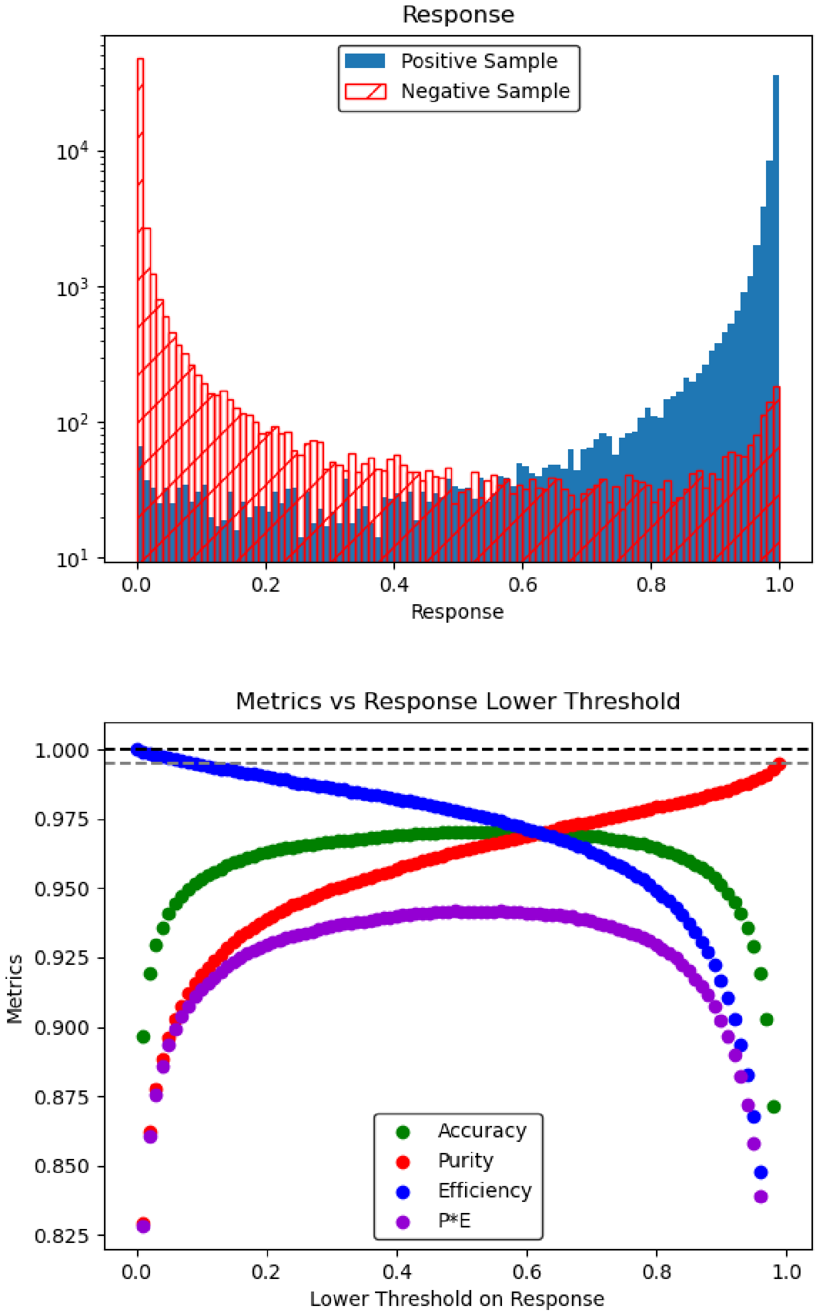


Figure 5.4: Top: The response of the learning algorithm forming the AI trigger. In blue the response for the events of the testing set with an electron, in red that for events without an electron. Bottom: The accuracy, purity, efficiency and the product of the purity and efficiency as a function of the threshold on the response. Plots taken from [122].

project in Appendix A.1 but hopefully it will here have served as a good example of why it is crucial to define metrics that are representative of the problem that one aims to solve so as to guide the decision of where to place a threshold on the response in classification problems. Note that the top histogram of Figure 5.4 shows a peak in the response of the positive sample at low response, and vice-versa for the negative sample. This is due to closely overlapping tracks that the model isn't able to separate leading to errors in the trigger prediction.

A final consideration around the choice of metrics is that, when using an algorithm trained on simulated data but applied to data taken with a real detector, these metrics become hard to calculate. Indeed, in the case of a PID problem, if the particle type was already known so as to evaluate what a True or False Positive or Negative is, then there would be no need for the learning algorithm in the first place. These metrics could be evaluated on the simulated data, but in that case they might not apply to the data taken with a real detector. Instead, one should look to other metrics that do not rely on the confusion matrix. For example, by fitting a resonance with some background, one can look at the signal to background ratio and choose to maximise it, or just the raw amount of signal and background, as a function of the threshold placed on the response.

The first, second and third step of the machine learning pipeline were first to produce the training data, second to train the algorithm and third to evaluate its performance on a testing dataset. The next step is to optimise what are known as the algorithm's hyperparameters. These are user defined components of the algorithm, as for example the choice of loss function. More detail will be given in the next section on what these hyperparameters are for different algorithms. Typically the second, third and fourth step will be repeated a few times until the performance of the algorithm is deemed satisfying, or as high as achievable. Finally, the last step of the machine learning pipeline is to deploy the algorithm on unseen data!

5.2 Algorithms Galore

There are many different kinds of supervised learning algorithms that have been developed over the last thirty years as machine learning has become more and more prevalent in modern computing practices. In fact, there are too many algorithms to describe here without completely derailing the narrative of this thesis. As such, emphasis will be placed on the algorithms used in later chapters of this thesis. In particular, these comprise decision trees and neural networks, and variations on these.

Decision trees (DT) [124, 125] are perhaps, conceptually, the simplest algorithms to describe. DTs, exemplified in Figure 5.5, are composed of nodes which branch out into child nodes. The last node at the end of a branch is known as a leaf and returns the predicted label. DTs attempt to split the training data by repeatedly splitting it at each node. The splits are performed by selecting simple cuts on a random choice of variables such that a specified loss function is minimised. One example of such a loss function is the Gini index G [124]:

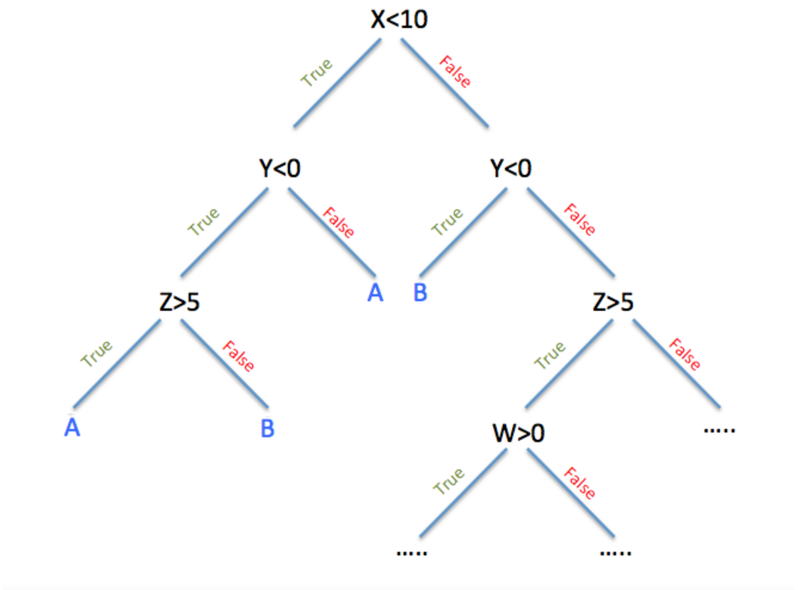


Figure 5.5: An example of a decision tree, with example cuts on variables W, X, Y, Z at each node, leading to a classification into classes A or B. Adapted from [123].

$$G = 1 - \Sigma_i \left(\frac{p_{mi}}{p_m} \right)^2 \tag{5.5}$$

where p_{mi} is the proportion of samples at node m belonging to class i. The splitting continues until each leaf contains one sample, or a specified maximum depth is attained, where the depth corresponds to the number of nodes in a branch. One common issue with decision trees is that these are prone to overfitting the training data [125]. For example, with an arbitrarily large maximum depth, the splitting can continue to produce a tree that will assign the right label to all entries in the training sample due to very specific cuts that can only correspond to the training sample and will not generalise to another sample.

One simple solution to reduce overfitting and generally improve the performance of decision trees is so called boosting. The idea behind boosting is that it is generally easier to train several smaller models than a single large model whilst still avoiding overfitting. Several different boosting algorithms, such as adaptive boosting (AdaBoost) [126] or gradient boosting (GradBoost) [127] will perform what is known as the boosting algorithm, but the overall set of steps required for boosting remains the same. Initially, the boosting algorithm will assign equal weights to each element of the training sample. A first DT is trained on the training sample and makes a prediction for each event. The boosting algorithm then assesses the model prediction with a given loss function, and increases the weight of elements of the training sample which have a larger error. A weight based on the DT's overall performance on the whole training sample is also assigned to the DT. The weighted data is then passed to the next DT, with the training influenced by the weight on each element of the training sample. As such, subsequent DTs will focus on elements that were poorly predicted by previous DTs. The boosting algorithm will then continue to train DTs, assign weights to the training data and pass the weighted training data to subsequent DTs until the training error reaches a certain threshold, or the number of DTs reaches a user defined maximum. The number of DTs can be loosely related to the number of epochs mentioned in the previous section, as each new DT being trained corresponds to a new optimisation process of the learning algorithm. At prediction time, the weighted mean of the prediction of each DT is then used as the final output. The weight used for each individual DT is the overall weight that was assigned to the DT based on its performance on the whole training sample. In such a way, boosting creates a single strong DT out of an ensemble of weak DTs, with the final algorithm generally referred to as a boosted decision tree (BDT).

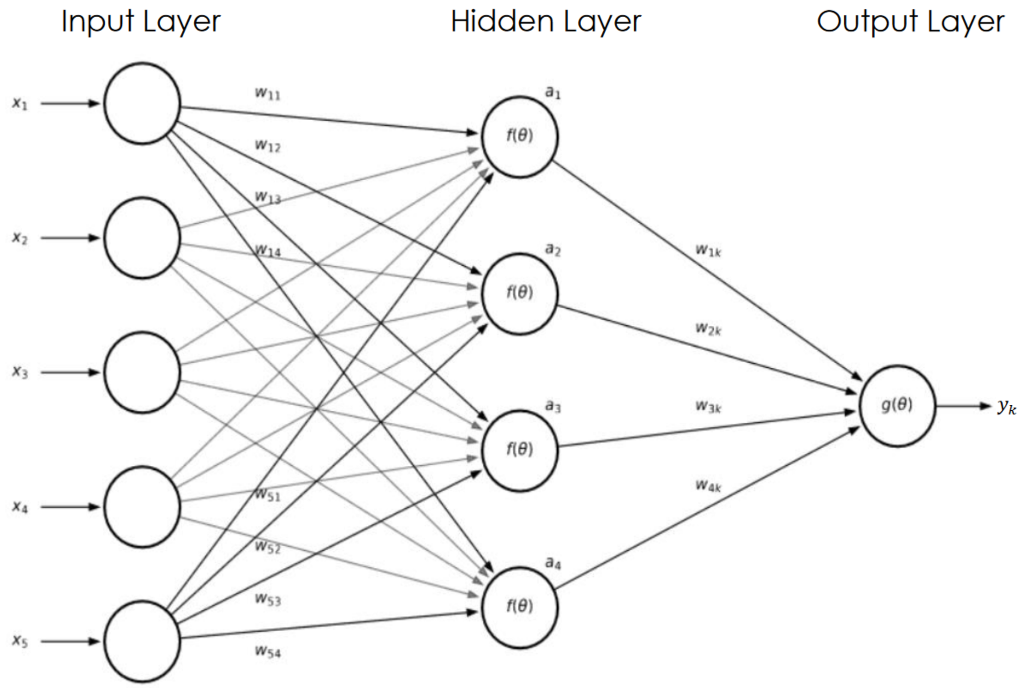


Figure 5.6: An example of a multilayer perceptron, with one hidden layer, five input nodes corresponding to five input features and one output node corresponding to one predicted label. Adapted from [128].

The other main class of learning algorithm used for the work presented in this thesis are so called neural networks (NNs) [129]. An example is given in Figure 5.6. NNs are composed of layers of nodes, often called neurons due to the origins of NNs based on creating learning algorithms that learn as a human brain does. Neurons are interconnected with the output of each neuron in a given layer passed onto each neuron in the subsequent layer. NNs will have an input layer which has as many neurons as there are features in the feature matrix **x** and an output layer with as many neurons as columns in the label matrix **y**. In the example given in Figure 5.6, each of the five columns of **x** is called x_i. The input of each neuron j in the hidden layer a_j is given by:

$$a_j = f(\Sigma_i(w_{ij}x_i + b_i)) \tag{5.6}$$

where w_{ij} is the weight of the neuron i in the input layer going to the neuron j in the hidden layer, b_i is the bias of the neuron i in the input layer and f is the activation function used in the hidden layer. The output of each neuron k in the output layer is then:

$$y_k = g(\Sigma_j(w_{jk}a_j + b_j)) \tag{5.7}$$

where the symbols have a similar meaning. The activation function is used to constrain the output of each neuron between zero and one. Typical examples are the tanh function or the ReLu function defined as $R(x) = max(0,x)$. Typically in binary classification tasks the activation function of the output layer is the sigmoid function to obtain the probability that the element of the training sample belongs to class k. For regression purposes the activation function of the output layer is often linear to avoid any transformation of the output of the last node which may make it harder for the NN to train. NNs are then trained by using variations of Stochastic Gradient Descent algorithms, so called optimiser algorithms, that aim to find the minima of the loss function as a function of the weights w_{ij} of each neuron in each layer of the NN. Popular choices of optimiser algorithms are ADAM and ADAMAX [130]. The weights are then changed using a backpropagation algorithm such that $\Delta w_{ij} = -\eta \frac{dL}{dw_{ij}}$ for L the loss function and η the learning rate which can be seen as a step size in weight space.

Figure 5.6 shows an example of a multilayer perceptron (MLP) [131] which are the most basic type of neural networks. Another type of NN is a convolutional neural network (CNN) [132, 133]. CNNs have become the go-to NNs for image classification and more generally computer vision. They are composed of convolutional layers that enhance or remove features of their input images, and a deep neural network which classifies the output of the convolutional layers. Appendix A.1 will describe how CNNs can be used to create a deep learning based electron trigger for CLAS12. Autoencoders [134] are a third type of NN, often used for image generation and denoising purposes. The idea is that a first network called an encoder will compress the feature space to a smaller latent space representation. A second network will then decode the latent space representation back to the feature space. This allows for the network to learn the important aspects of the feature space which are encoded into the latent space. By removing the encoder, random entries can be passed to the decoder perpetrating to mimic the latent space representation, leading to the generation of entries in feature space which resemble what the autoencoder was trained on. Autoencoders often combine convolutional layers with a typical neural network and are then called convolutional autoencoders.

As mentioned at the end of the previous section, for all types of learning algorithms, some parameters of the algorithm which are not optimised by the neural network are left to be defined by the user. These are typically referred to as hyperparameters. For BDTs, these can be for example the number of trees, the maximum depth of the trees, the loss function, the boosting algorithm or the minimum number of training events required to pass through each node. For NNs, the hyperparameters are for example the number of nodes, the number of neurons in each node, the optimiser, the loss function in each layer, the learning rate, or the number of training epochs. What hyperparameters are available for tuning will typically depend on the software used to implement the learning algorithm. In python, a very common software package to train and deploy NNs is *tensorflow* [135], with the *scikit-learn* [136] package being very useful to train and deploy DTs and MLPs. In Java, the *DeepLearning4J* [137] package can train and deploy NNs, including NNs trained with *tensorflow*. In C++, the *ROOT TMVA* [138, 139] package allows to train and deploy DTs, MLPs and NNs.

To end this chapter, a final mention is given to the type of learning algorithm used in the accompanying denoising tutorial. As mentioned above, autoencoders are well suited for this kind of task, as they take as input the noisy image and return the de-noised image, after mapping the input to a lower dimensional latent space representation. The suggested autoencoder in the online tutorial is very simple with an input layer taking an array of size $S \times W$, a second layer that flattens the two dimensional array into a one dimensional array, three hidden layers that decrease then increase in size and finally an output layer which reshapes the one dimensional output of the hidden layers into a two dimensional array of size $S \times W$. The first two hidden layers have a ReLU activation, with the last hidden layer having a linear activation. This helps the network produce a sensible output which resembles the single track in the drift chambers. The optimiser here is ADAMAX with a mean squared error loss function. The network is trained using *tensorflow* in python.

Chapter 6

Final State Particles

Once the data taken by CLAS12 has been reconstructed and calibrated, it is then ready for analysis. Typically the analysis of data taken by a high energy physics experiment starts with particle identification (PID). To be able to observe a given reaction, all or a subset of the final state particles produced in the reaction need to be detected. These particles then need to be identified by its particle type, to make sure that the right decay products are being detected. Finally, once detected and identified, corrections typically need to be made to the reconstruction of these particles. For example, some issues can arise in measuring the momentum of the final state particles. This chapter will discuss the detection, identification and corrections applied to the different particle types that were required for the analysis of J/ψ photoproduction at CLAS12. As such this chapter marks the start of the actual analysis work that went towards the measurements proposed in Chapter 2. A final consideration is that seeing how PID procedures affect the number of measured J/ψ (here called J/ψ yield), and the number of background events, is necessary to estimate the impact of these procedures. Chapter 7 will detail further analysis procedures required to measure these J/ψ and background yields. As such, before reading Chapter 7, the J/ψ yield should be taken as a metric to maximise whilst the background yield should be taken as a metric to minimise, which allows to evaluate certain PID procedures. Finally, most of the analysis procedures described here have been incorporated into the *chanser* [112] framework to make it easier to deploy these during analysis. More detail is given in Appendix A.2.

6.1 Protons and Charged Hadrons

For the analysis of J/ψ photoproduction at CLAS12, the simplest particle to detect and identify is the proton. As discussed in Chapter 3, protons are detected based on hits in the DC and the FTOF. The momentum and charge of the proton is measured by the DC, and used in conjunction with the velocity measured by the FTOF to identify protons and distinguish them from other charged hadrons as shown in Figure 3.7. As detailed in Chapter 3, a general particle identification scheme is enforced by the event builder (EB). This particle identification scheme is named EBPID. For protons, and charged hadrons in general, the EBPID is based on the difference, corrected by the start-time t_{st}, between the time t_{FTOF} recorded by the FTOF and the expected time based on the path length L and velocity $\beta_h(p)$ for a given hadron mass hypothesis at a given momentum [140]. This results in a Δt calculated as:

$$\Delta t = (t_{FTOF} - t_{st}) - \frac{L}{c\beta_h(p)}; h = d/p/K^{\pm}/\pi^{\pm} \tag{6.1}$$

where:

$$\beta_h(p) = \frac{p}{\sqrt{p^2 + m_h^2}} \tag{6.2}$$

for p the momentum of the particle and m_h the mass hypothesis of the particle. The particle is then assigned the particle type for the mass hypothesis that minimizes Δt, which corresponds to the best hypothesis for the particle. The distance from the expected vertex time can then be evaluated, leading to a so called chi2pid value [140]. Cutting on this value results in narrow β versus momentum bands, with a cut at three sigma allowing to tighten the particle identification requirements for charged hadrons [140]. From now on, any particle called a proton, kaon, pion or deuteron at CLAS12 can be taken as a particle identified as that given particle type by the EBPID. Figure 6.1 shows the typical momentum distribution of protons in J/ψ photoproduction data at CLAS12. This data is taken in the right kinematic ranges for J/ψ photoproduction but contains many more events than just actual J/ψ events. Comparing this histogram to the rightmost graph of Figure 3.7, it is clear that the typical momentum range of protons for J/ψ photoproduction places these in a range where they are easily identified by the event builder. No further refinements on proton identification were considered.

Figure 6.2 shows the momentum and polar (θ) and azimuthal (ϕ) angles resolution of protons detected in the FD. These were obtained in simulated data, created as detailed in Chapter 4, by taking the difference of the true generated four momentum of the protons to that reconstructed from the simulated data. As shown, the resolution distributions of the polar and azimuthal angles are centred on zero, and the momentum resolution is only offset by 5 MeV, so no offset corrections to the four-momentum components are necessary. The width of the resolution distributions are comparable to the summary of the CLAS12 performance given in Table 3.1.

Finally fiducial cuts can be applied to tracks on the edge of the drift chambers. These cuts are determined by fitting the χ^2 NDF value of tracks, which quantifies the track quality, as a function of the angular or cartesian coordinates of hits in the calorimeters [140]. Typically the track quality degrades close to the edge of the detector which leads to poorly reconstructed tracks. Figure 6.3 shows the effect of the fiducial cuts for positive pions and protons. Fiducial cuts allow to ensure only well reconstructed tracks are used in the analysis, which in this case

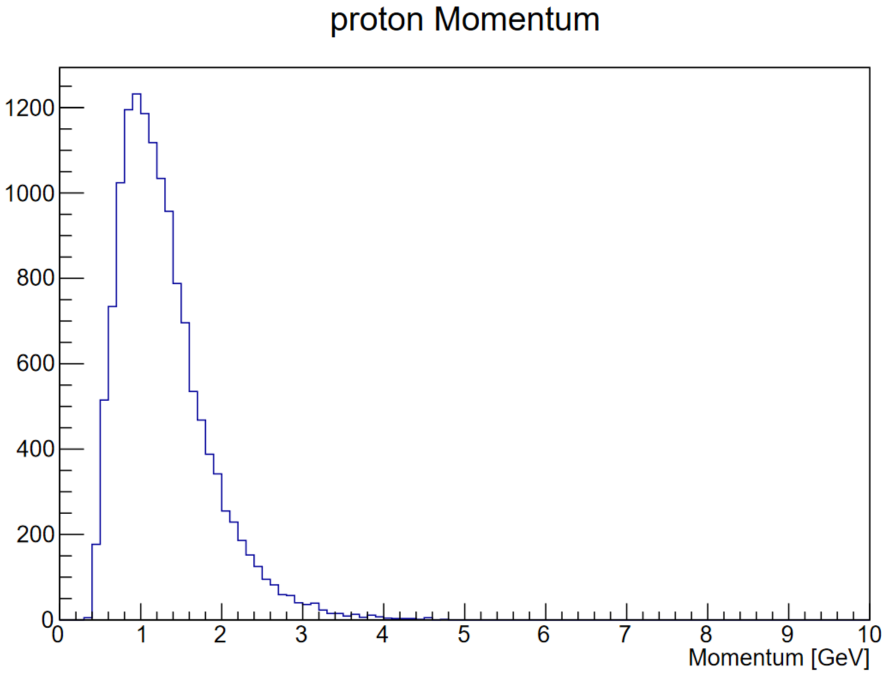


Figure 6.1: The proton momentum in J/ψ photoproduction data taken by CLAS12.

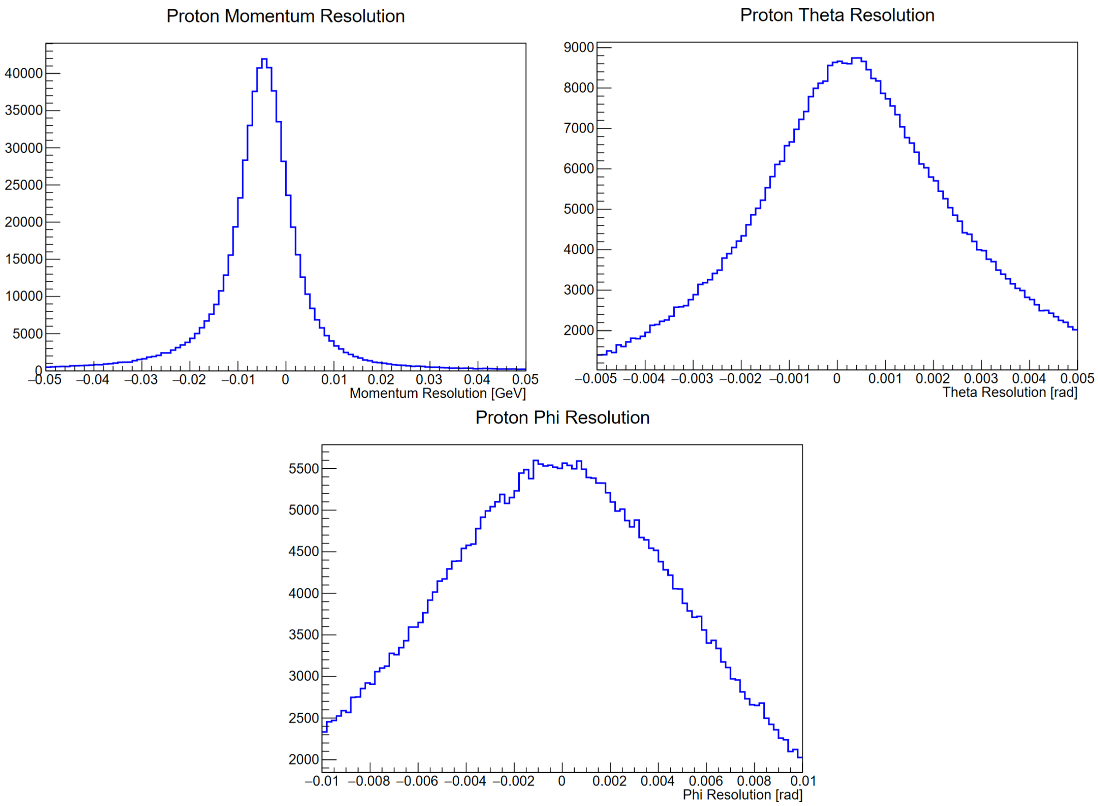


Figure 6.2: The proton momentum, θ and ϕ resolution modeled using Monte Carlo simulations.

ensures good measurements of momentum. This is especially important as simulated data may not be able to recreate poorly reconstructed tracks.

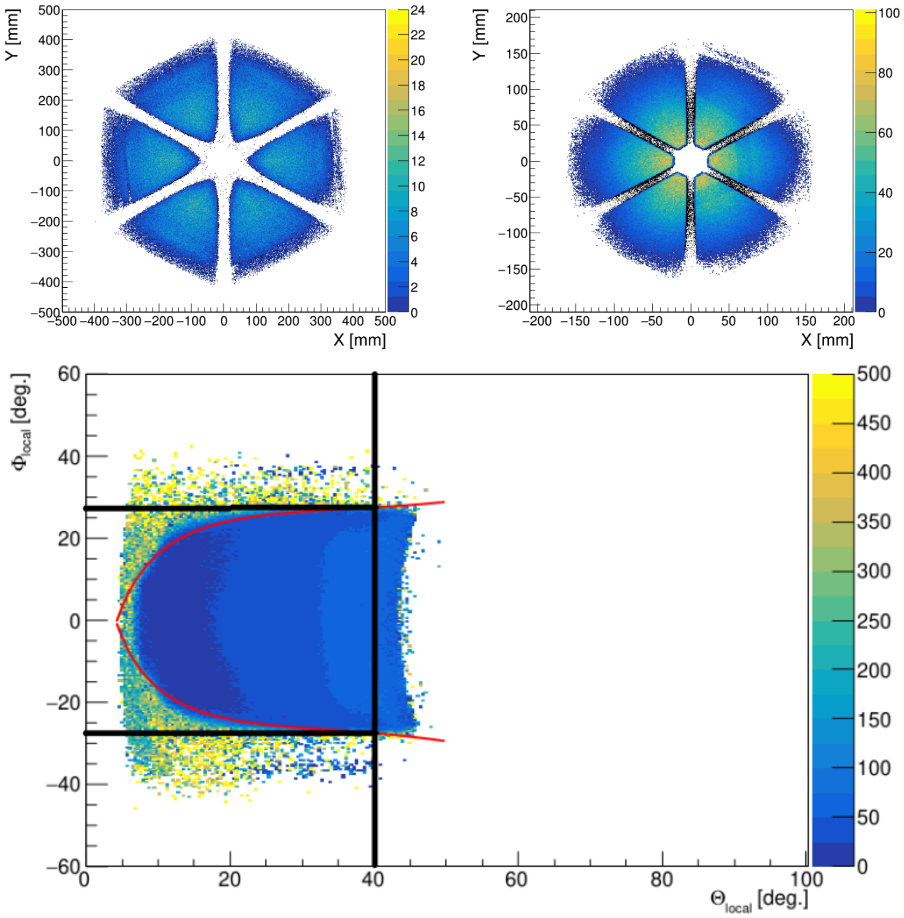


Figure 6.3: Top Row The effect of the fiducial cuts in X and Y coordinates of the drift chamber for positively charged pions. The right most plot has all hits whilst the left most plot shows the effect of the fiducial cuts. Bottom: The angular fiducial cuts (red) for protons. The intersection of the black lines denotes a fixed point for the fit to negate the effect of fluctuations in the track quality assessment at high values of θ. All three images taken from [140].

6.2 Muons

Muons are minimum ionising particles (mips), meaning that these will have low energy deposition in the calorimeters. Muons are therefore detected using the DC to measure their momentum and charge and the calorimeters to measure their energy deposition. No EBPID assignment is made for muons, due to the fact that they have a very similar mass to pions, and the hadron particle identification scheme described in the previous section would not be able to distinguish between muons and pions. As such all muons are identified as pions and sometimes kaons by the event builder.

The first step to identifying muons is to place a cut on their energy deposition in the calorimeters. Figure 6.4 shows the energy deposition taken from simulated data for the positive and negative muons in each of the PCAL, ECIN, ECOUT. Overlayed are the energy deposition of simulated pions and that of muon candidates in J/ψ photoproduction data from CLAS12. As can be seen, pions are also mips and their energy deposition distributions peak roughly at the same values as muons. However, due to hadron-hadron scattering in the calorimeters, pions can produce showers leading to tails in their energy depositions. A similar behaviour can be observed for kaons.

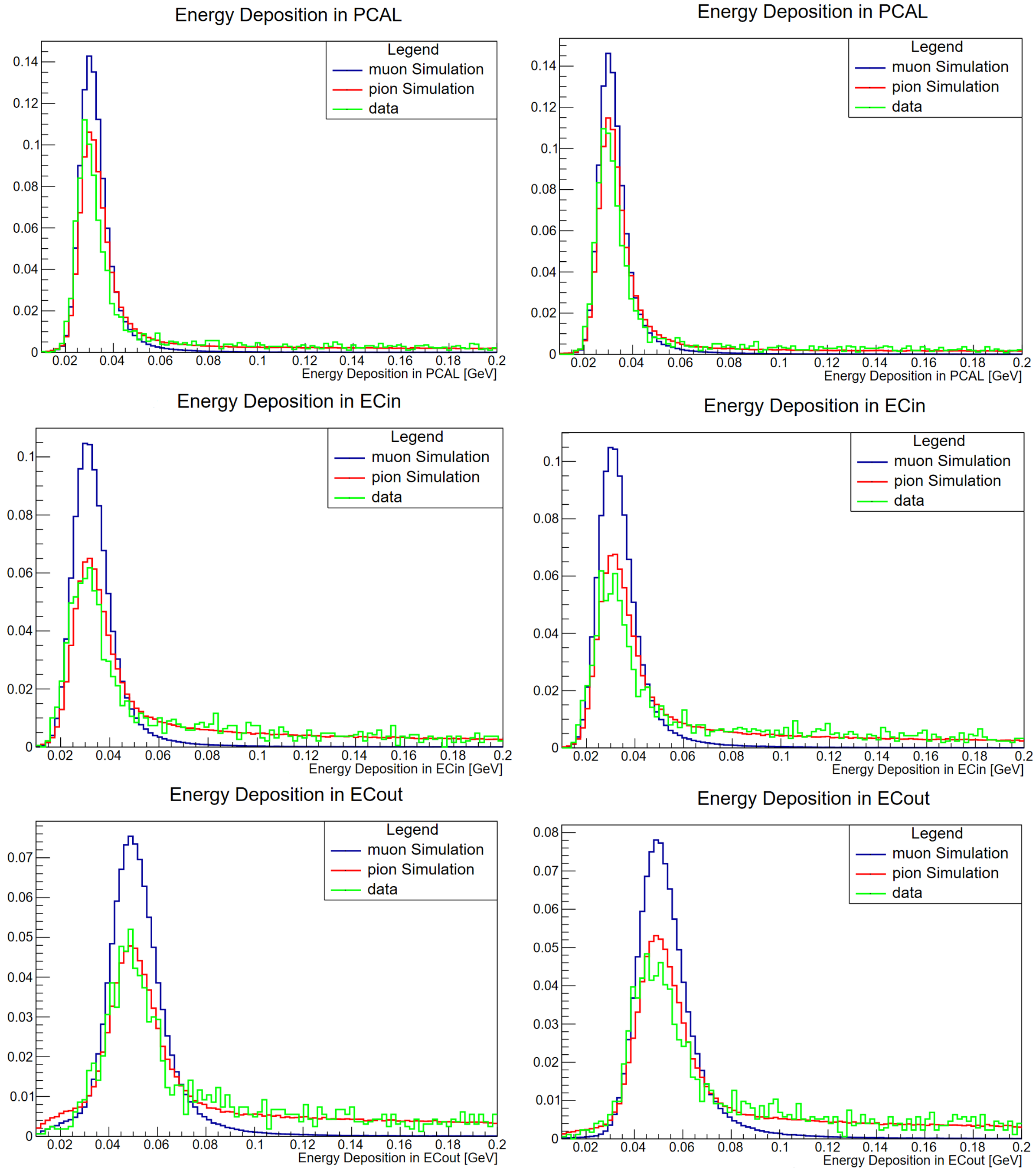


Figure 6.4: The energy deposition of simulated negative (left) and positive (right) muons in each of the PCAL (top row), ECIN (middle row) and ECOUT (bottom row). Overlayed are the energy deposition of simulated pions, and that of muon candidates in J/ψ photoproduction data from CLAS12. All histograms are scaled by one over their integral so that the sum over all bins returns one.

As J/ψ does not decay to a pion pair and has a high mass (relative to the typical beam energies at CLAS12), the di-pion background does not completely overwhelm the J/ψ peak, meaning that the peak is still clearly visible. An example of the di-muon invariant mass is shown in Figure 6.5. The J/ψ peak is fitted with a gaussian function whilst the background is fitted with a third order polynomial. A parameter of the gaussian fit function returns the J/ψ yield, whilst the background yield is taken as the integral of the background function between 2.98 and 3.2 GeV. These also require some cuts on the missing mass squared and Q^2 as described in Chapter 2, with more detail on the effect of the cuts given in Chapter 7.

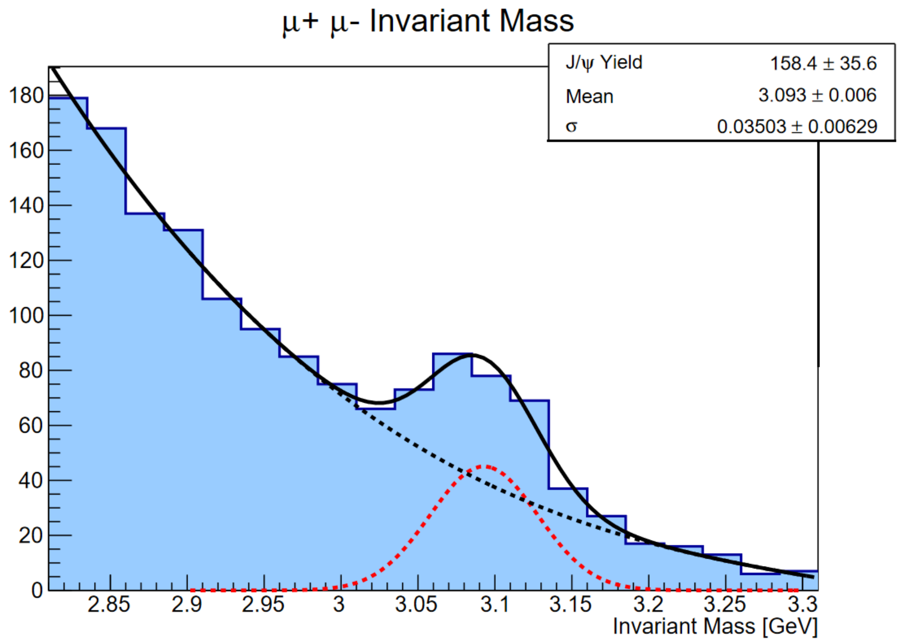


Figure 6.5: An example of the di-muon invariant mass for quasi-real photon energies of 8.2 to 10.6 GeV. The J/ψ peak is clearly visible and fitted with a gaussian function. The parameters of the gaussian fit function are indicated in the top right box. One of the parameters returns the J/ψ yield. The background is fitted with a third order polynomial.

The data used for the J/ψ photoproduction analysis as described in Chapters 2 and 4 matches up well with the simulated pion distribution, as shown in Figure 6.4, suggesting that most muon candidates are in fact pions when considering muon candidates over a large invariant mass range, instead of around the J/ψ mass. As mentioned in Chapter 3 the di-muon trigger already enforces energy deposition cuts. For the PCAL the energy deposition is required to be between 15 and 60 MeV. For the ECIN and ECOUT, the sum of their energy deposition is required to be between 60 and 120 MeV. This is to minimise the number of misidentified pions that fire the trigger. Figure 6.6 shows how tightening these requirements by lowering the upper limit on the energy deposition affects the J/ψ and background yields for each of the PCAL, ECIN, ECOUT. The energy deposition cuts are applied to both the positive and negative muon simultaneously. Figure 6.6 informed the decision to maintain the upper limit on energy deposition at 60 MeV for the

PCAL and set this at 110 MeV for the ECOUT, whilst the upper limit on the energy deposition in the ECIN was set to 80 MeV.

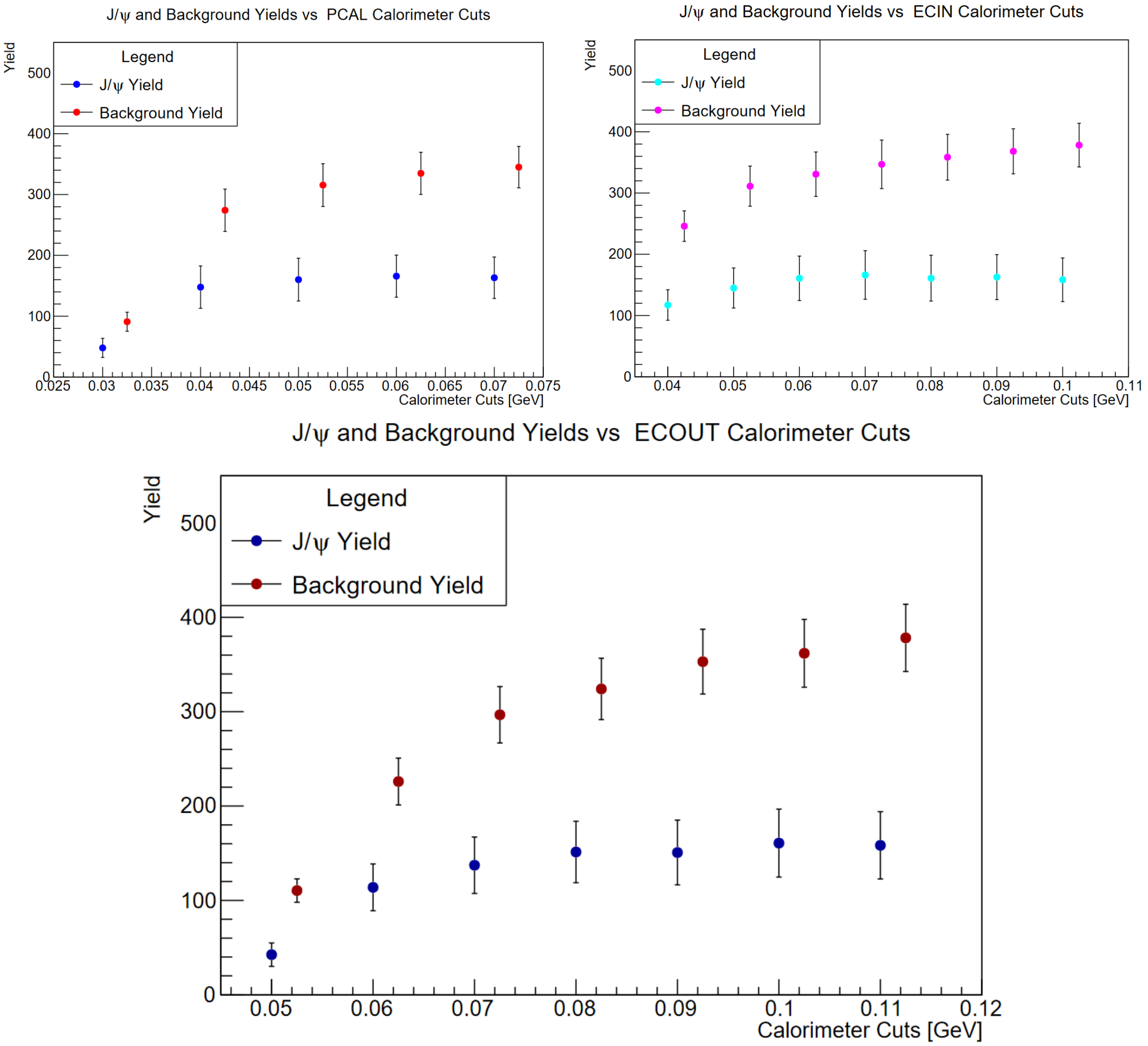


Figure 6.6: The J/ψ and background yields as a function of the cut applied to the PCAL (top left), ECIN (top right) and ECOUT (bottom) energy deposition. The energy deposition cuts are applied to both the positive and negative muon simultaneously. Note that the background yield was evaluated at the same cut value as the J/ψ yield but was displaced by 2.5 MeV so that the error bars do not overlap.

As shown in Figure 6.6, decreasing the upper limit on energy deposition decreases the J/ψ yield and the background yield without substantially decreasing the J/ψ to background yield ratio. This makes sense given the plots of Figure 6.4; both pions and muons have a similar peak in their energy deposition and cutting away the upper edge of that distribution lowers the rates of muons and pions in a similar way. Simple cuts on the energy deposition are therefore not sufficient to improve muon identification at CLAS12.

Instead, given that machine learning is typically well suited to classification tasks such as particle identification, muon identification was refined using a boosted decision tree (as described in Chapter 5). This was trained and deployed using the *chanser* implementation of *ROOT TMVA*. The training and testing samples were created using simulated data. In this case, the decision was made to identify both positively and negatively charged muons at the same time. The logic here is that both would be required for the analysis of J/ψ photoproduction and both are poorly identified by the CLAS12 reconstruction algorithm. The training data set therefore contained features for the simulated muon pair as the positive sample and a simulated pion pair for the negative sample. To avoid any bias towards the J/ψ mass, the simulations were made for phase space photoproduction of two pions or two muons such that the invariant mass presented a decaying slope from 2.5 GeV onward, with no J/ψ peak present in the muon simulation. This also ensured that the four-momentum distributions of the muon or pion pair were the same so as to avoid any bias towards four-momentum.

The features used for classification for both the positive and negative particles first contained information taken from the calorimeter such as the energy deposition in each of the PCAL/ECIN/ECOUT, the total sampling fraction defined as the sum of the energy deposited in each calorimeter divided by the momentum of the particle, and the sampling fraction in each of the individual calorimeters. The number of strips in each of the three views that contained a hit in the PCAL was also used as pions tend to produce showers and therefore have a higher probability of firing several strips than muons. The distance of the peak in each of the three views of the cluster from the edge of the PCAL was included, the logic here being that a hit close to the edge might fire fewer strips which should be accounted for. The second moment in each of the three views of the calorimeter are also included. The number of photoelectrons produced in the HTCC was added to the set of features, as pions are known to fire the HTCC at high momentum, and the energy deposited in the FTOF was also added as this can relate to the Bethe Bloch equation. Finally, the features used to identify muons from pions included the four momentum components of the particles. The logic here is that as the distributions were similar between the positive and negative samples this would not lead to any bias but would allow to correlate to momentum dependent quantities, such as the sampling fraction.

The decision to use boosted decision trees for muon identification was based on the fact that decision trees are not negatively impacted by the inclusion of unnecessary variables. The cut imposed on an unnecessary feature in a node of the decision tree can be set such that the cut does not negatively impact the performance of the tree. This is useful in cases such as these where the optimal set of features is not known, and additional features can be included without too much concern for the performance of the learning algorithm. The default *ROOT TMVA* hyperparameters were used for the boosted decision tree. Some work went into optimising these but this lead to small gains on the simulated testing dataset which did not necessarily translate to the performance of the decision tree on CLAS12 data.

Figure 6.7 shows the boosted decision tree response when applied to the simulated training and testing sets, after separating these into the positive and negative samples that correspond to the di-muon and di-pion samples. Note that here the response does not range from 0 to 1 due to the specificities of the implementation of the decision trees in *ROOT TMVA*. The response for the testing set is given by a histogram whilst the response for the training set is given by dots. The fact that these overlay well denotes that the boosted decision tree is not overfitting. A Kolmogorov-Smirnov test between the response on the training and testing set also returns a low value which again suggests that the decision tree is not overfitting. Figure 6.7 also shows the J/ψ and background yield as a function of the lower limit threshold on the response. The yields were evaluated on J/ψ photoproduction data taken with CLAS12 and not simulated data. The two plots agree with one another in the sense that the background and J/ψ yields drop off close to where the negative and positive responses start to increase, with the exception of a secondary peak in the negative response from -0.4 to -0.2. This is due to pions with energy deposition outside of the trigger cuts, which were kept in the training sample in case of a pion contamination due to events selected by another trigger than the di-muon trigger. The agreement between the two plots of Figure 6.7 suggests that the simulated data is representative of CLAS12 data, except for some background events that are not present in the CLAS12 data when enforcing the trigger requirements. The boosted decision tree is able to successfully distinguish muons from pions in CLAS12 data, as demonstrated by the fact that the background yield drops off faster than the signal yield. This then leads to better muon identification.

A visual test of the assertion at the end of the last paragraph is to plot the di-muon invariant mass at different values of the cut on the response. As shown in Figure 6.8 the background under the J/ψ peak decreases as the lower threshold cut on the response increases. An important consideration is here the gaussian fit parameters. An increase in sigma as the cut on the response increases would indicate a bias towards the J/ψ mass, essentially indicating that the fit hypothesis is modified by the boosted decision tree and artificially increasing the measured J/ψ yield. As this is not the case, the boosted decision tree based di-muon identification is unbiased and offers a clean approach to muon identification. This amounts to the first muon identification algorithm at CLAS12. This is the first analysis to explicitly observe and identify muons with CLAS12. From now on, any particle called a muon can be taken to be a mip candidate with a cut on the di-muon response of at least -0.12, which is where the J/ψ yield starts to drop.

A final consideration from Figure 6.8 is that the mean of the gaussian fit to the J/ψ peak is very close to that of the J/ψ mass of 3.097 GeV. This suggests that the four momentum components of the di-muon pair are well reconstructed, as otherwise the peak would be shifted

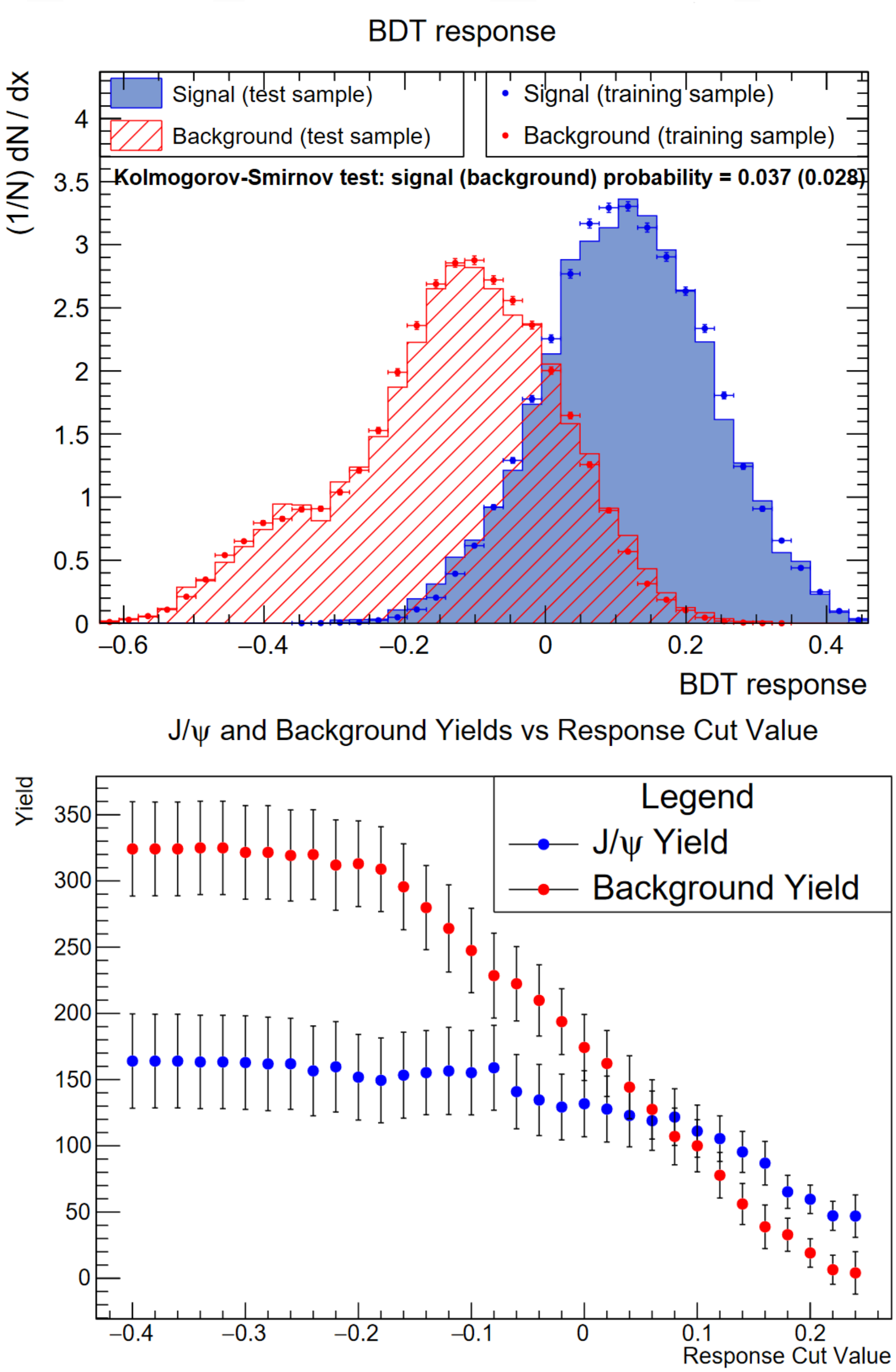


Figure 6.7: Top: The boosted decision tree (BDT) response when applied to the simulated training and testing sets, separated into the di-muon (blue) and di-pion (red) samples. Bottom: The J/ψ and background yield as a function of the lower limit cut on the response. The yields were evaluated on J/ψ photoproduction data taken with CLAS12 and not simulated data. Note that the background yields are slightly decreased when compared to Figure 6.6 as an additional cut was placed on the ECIN energy deposition.

further away from the J/ψ mass. No further corrections were applied to the di-muon candidates. Finally, fiducial cuts to remove hits close to the edge of the drift chambers can be applied, as described in the previous section for charged hadrons. These have very little effect on the J/ψ

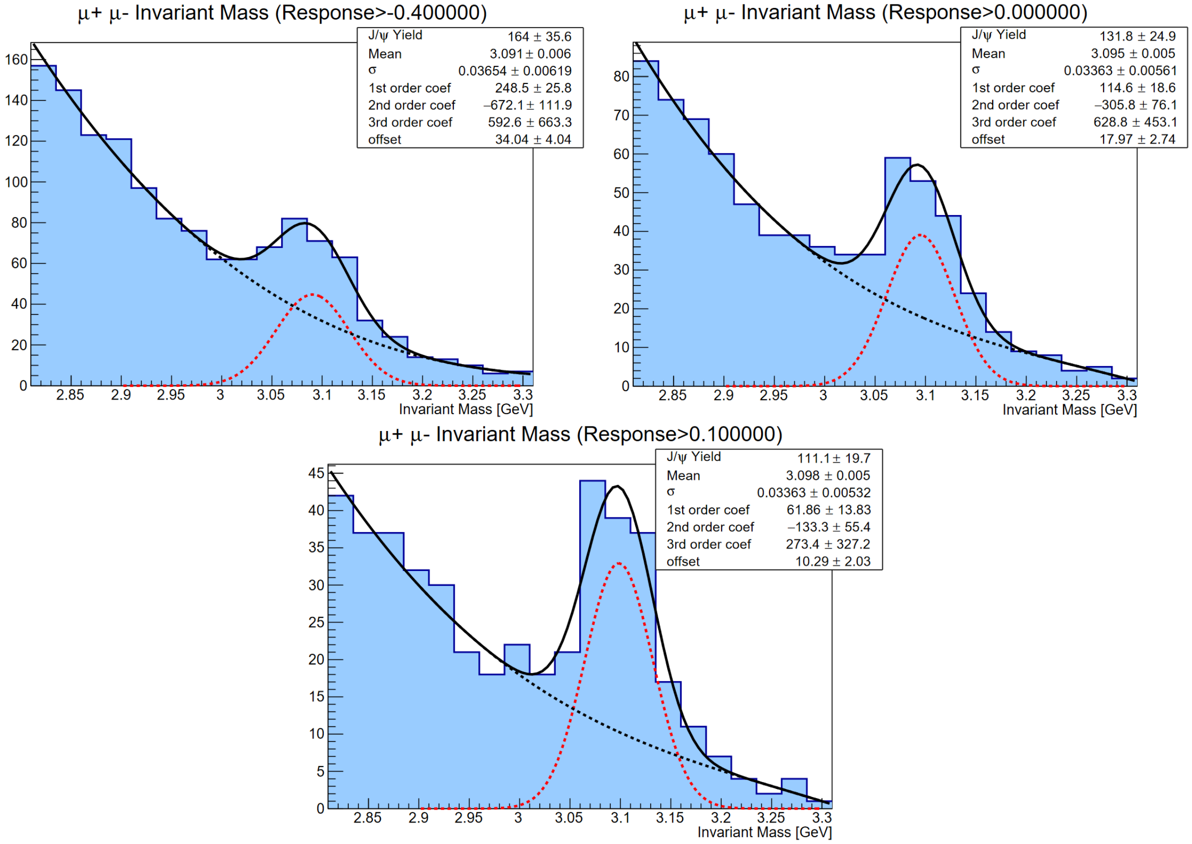


Figure 6.8: The di-muon invariant mass for a cut on the response at -0.4 (top left), 0 (top right) and 0.1 (bottom). A cut on the response of -0.4 can be taken as not using the boosted decision tree. The J/ψ peak is fitted with a gaussian, whilst the background is fitted with a third order polynomial. The J/ψ yield decreases as the cut on the response is increased, as does the background. The gaussian fit parameters are consistent within errors for different cut values on the response which suggests that the decision tree is not biased towards the J/ψ mass.

yield, whilst removing a marginal amount of background, but help ensure well reconstructed muon candidates.

6.3 Electrons and Positrons

As detailed in Chapter 3, electrons and positrons are detected with a conjunction of a track in the DC, energy deposition in the calorimeters and a signal in the HTCC. Figure 3.6 exemplifies how electrons produce showers in the calorimeters which lead to a high sampling fraction of about a quarter. A parametrisation is established for both the mean and standard deviation of a normal distribution fitted to slices of the sampling fraction in bins of momentum. The EBPID for electron and positrons then identifies these by requiring two photoelectrons in the HTCC matched to a positive or negative track in the DC, with an energy deposition in the PCAL above 60 MeV. The sampling fraction for electrons and positrons is then required to be within five standard deviations from the mean for a given momentum. Note that muons and electrons have opposite requirements on the energy deposition in the PCAL, ensuring that neither particles types are misidentified as one another which avoids any interference between the analyses of J/ψ decaying to a di-electron or di-muon pair. From now on electrons can be taken as electrons or positrons identified by the event builder.

However, as mentioned in Chapter 3, pions with momentum greater than 4.5 GeV can fire the HTCC, meaning that these will produce at least two photoelectrons. As shown in the previous section, pions can also have a high energy deposition due to hadron-hadron scattering which produces tails in their energy deposition distributions. As such pions can be misidentified as electrons. Figure 6.9 shows the electron and positron sampling fraction as a function of the total deposited energy, and the PCAL sampling fraction as a function of the ECIN sampling fraction. For both the electron and positrons, although in particular for positrons, some structure is clearly seen at low sampling fractions. This corresponds to pions misidentified as electrons or positrons by the event builder. A first refinement on the EBPID is to tighten the cut on the sampling fraction to three standard deviations from the mean parametrisation of the sampling fraction. Similarly the minimum energy deposition requirement can be increased to 70 MeV [140]. Another cut, so called triangular cut denoted by the red line in the bottom plots of Figure 6.9, is to require the sampling fractions in the ECIN and PCAL as [140]:

$$\frac{E_{ECin}}{P} < 0.2 - \frac{E_{PCAL}}{P} \tag{6.3}$$

Another fiducial cut requires the electron and positron hits in the PCAL to be at a distance of at least nine centimeters from the edge of the calorimeters [140]. This is due to the fact that electron showers may not be fully contained for hits close to the edges which leads to a decreased sampling fraction, as shown in Figure 6.10. This decreased sampling fraction will then affect the electron or positron PID. As such it simplifies analyses requiring electrons or positrons to remove particles close to the edge of the PCAL. The cut was placed at a distance of two strips which equals nine centimeters as each strip is 4.5cm wide. The cut is only applied in the PCAL, as a hit close to the edge of the PCAL will be close to the edge of the ECIN and ECOUT.

The effect of each cut on the sampling fraction, energy deposition and distance from the edge on the J/ψ and background yield for J/ψ decaying to an electron positron pair is shown in Figure 6.11. These were measured as before by fitting the electron positron invariant mass with a gaussian for the J/ψ signal and a third order polynomial for the background. The two most effective cuts are tightening the sampling fraction for electrons and positrons, and the triangular cut for positrons. As shown in Figure 6.9 there seems to be more pion contamination for positrons which leads to a greater decrease in background when applying the cuts of Figure 6.11. This is thought to be due to two pion production $ep \rightarrow e'\pi^{+}(\pi^{-})p$ where the negative pion goes

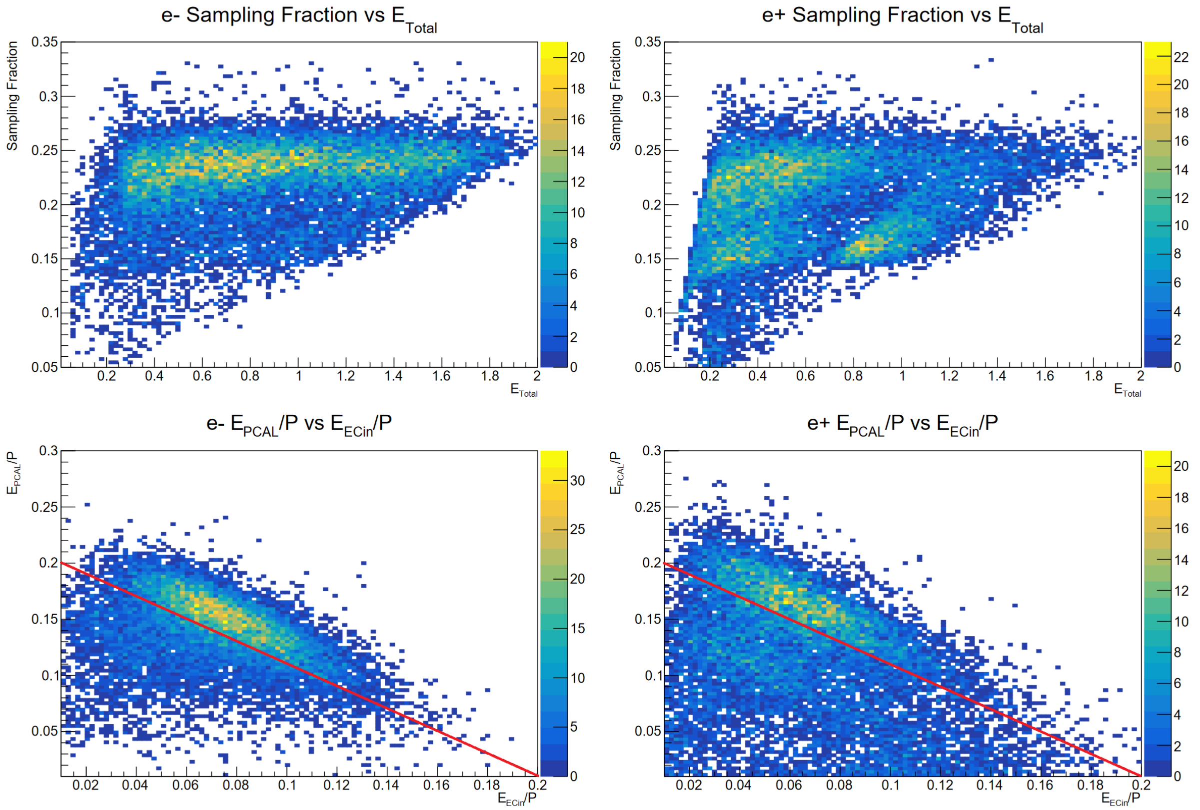


Figure 6.9: Top: The electron (left) and positron (right) sampling fraction as a function of the sum of the energy deposited in the PCAL, ECIN and ECOUT. Bottom: The electron (left) and positron (right) PCAL sampling fraction as a function of the ECIN sampling fraction. A suggested cut is denoted by the red line aiming to select particles above the red line.

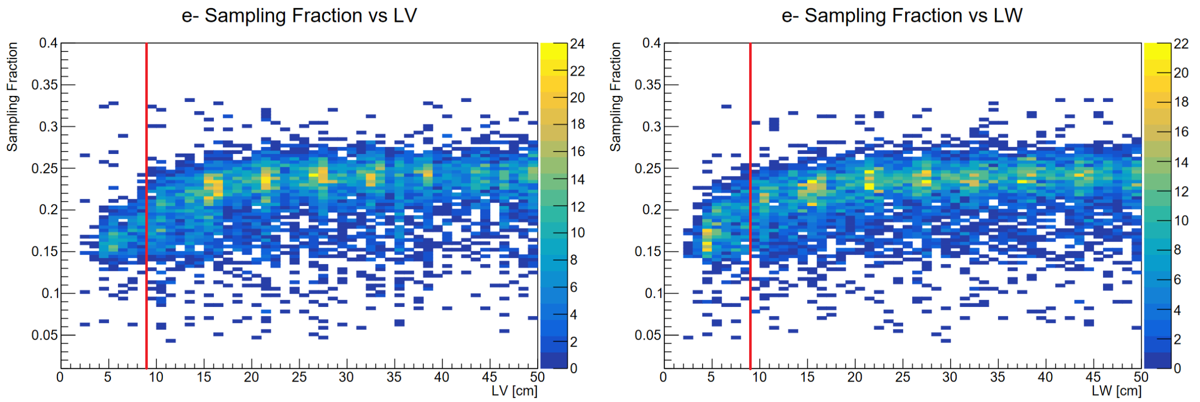


Figure 6.10: The electron sampling fraction as a function of the distance from the edge of the V view (left) and W view (right). LV/LW denotes this distance from the edge. The red line denotes the suggested cut at 9cm.

undetected, the positive pion is misidentified as a positron and the scattered electron is correctly identified as an electron. This then leads to a more sizeable background than misidentifying both pions which is less probable.

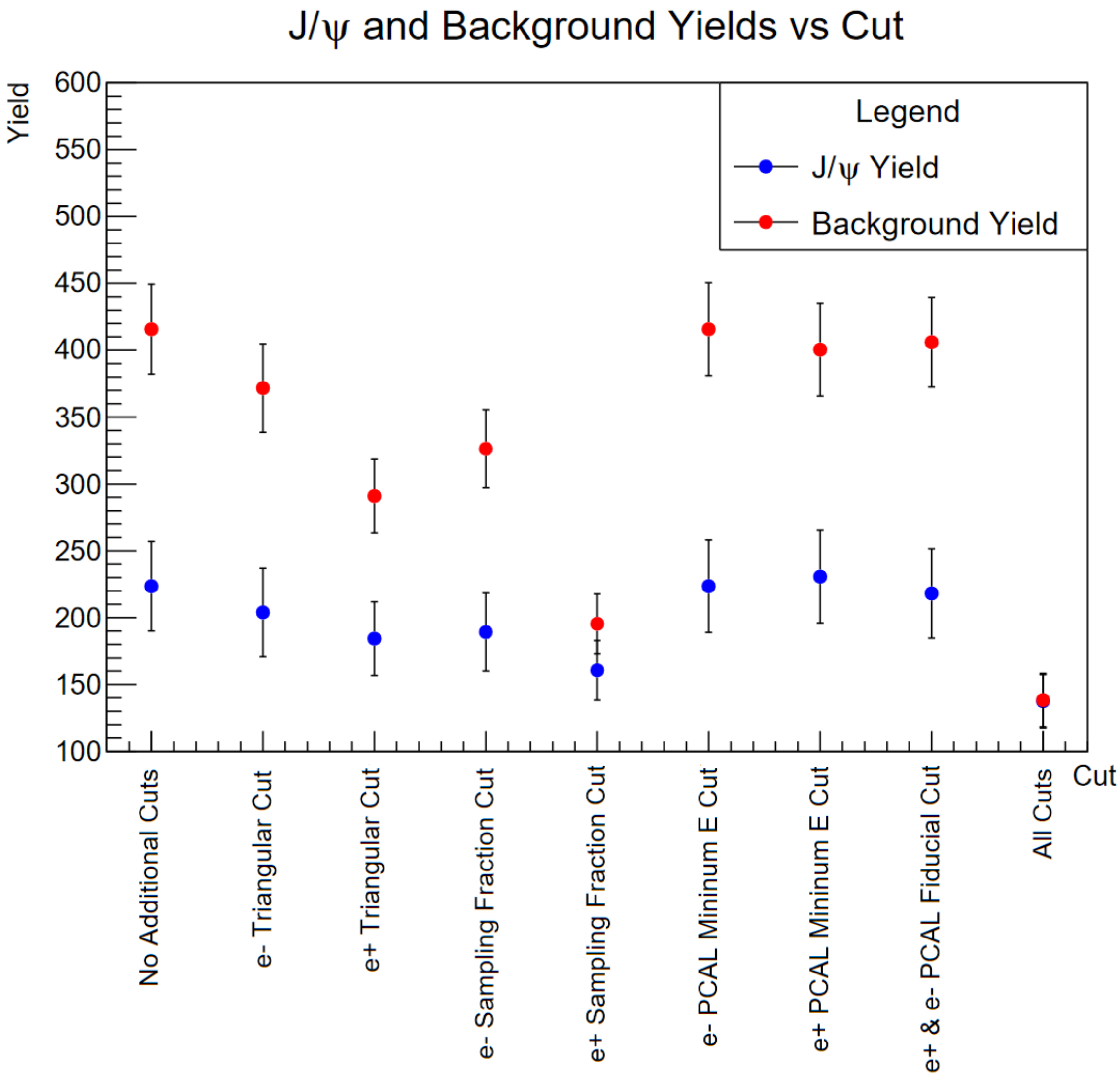


Figure 6.11: The J/ψ and background yield as a function of the cut applied to the analysis. Here the background is measured from 2.9 to 3.3 GeV in the e^+e^- invariant mass.

However, as shown, adding all the cuts does significantly reduce the J/ψ yield. The choice was therefore made to refine the positron PID using machine learning. Similar approaches were taken by CLAS12 collaborators when studying J/ψ photoproduction on a hydrogen target in the J/ψ decay to a positron and electron pair [84] or Timelike Compton Scattering at CLAS12 [141, 142]. These used different types of algorithms and training samples. A slightly different approach will be described here which is more in line with the muon identification procedure described in the previous section.

As before, the machine learning based positron PID used the default boosted decision tree architecture provided by *ROOT TMVA*. Here the training dataset was made from one simulated positron for the positive sample and a simulated positively charged pion as the negative sample. These were taken from a simulation of a phase space pair of oppositely charged electrons and pions made as before without any resonances at the J/ψ mass. The positron and pion therefore have similar kinematics, except below roughly 4.5 GeV in momentum, as the pion was further required to have been identified as a positron by the event builder. Very few pions are misidentified below 4.5 GeV, as they do not fire the HTCC. The full momentum range was allowed for both particles. This allows the algorithm to learn that it should only predict a particle as a pion above 4.5 GeV.

The variables used for training are similar to those used for muon identification. These are the energy deposited in each of the ECIN, ECOUT and PCAL, with the individual sampling fraction. The training dataset further contains the second moments in each view of the PCAL, and the distance from the edge of the calorimeter in each view as this can affect the sampling fraction. The number of photoelectrons produced in the HTCC is also included.

Figure 6.12 shows the boosted decision tree response when applied to the simulated training and testing sets, after separating these into the positive and negative samples that correspond to the positron and pion. As before, the response for the testing set is given by a histogram whilst the response for the training set is given by dots; both overlay well denoting that the boosted decision tree is not overfitting. Figure 6.12 also shows the J/ψ and background yield as a function of the lower limit threshold on the response. The yields were evaluated on data taken with CLAS12 used for the measurement of J/ψ photoproduction and not simulated data. The top and bottom panels of Figure 6.12 agree less well with one another than the respective plots shown for muon identification. The background and J/ψ yields drop off close to where the negative and positive responses start to increase. Above 0.4, the background yield decreases at the same rate as the J/ψ yield as the cut on the response is now equally rejecting sources of background with a true positron and positrons from J/ψ photoproduction. However, the constant decrease in background yield from -0.2 to 0.4 where the response on simulated data has very few pions suggests some difference between the training data and CLAS12 data. This may be due to larger tails in the energy deposition distribution for pions. The pions at the tails would be harder to distinguish from positrons, placing them at a higher response values. However the opposite effect would happen when distinguishing between muons and pions. Larger tails in the energy deposition would make the pions in the tails easier to distinguish from muons, explaining why the same issue was not seen for the muon identification algorithm. However, the fact that the background yield drops off at a lower response than the signal yield demonstrates that the boosted decision tree is able to distinguish between positrons and pions, which then leads to cleaner positron identification.

At a response cut of 0.3 almost a third of the background is removed, whilst most of the J/ψ signal is retained. Figure 6.13 shows plots of the sampling fraction as a function of the total energy deposited and of the sampling fraction of the PCAL versus that of the ECIN, once

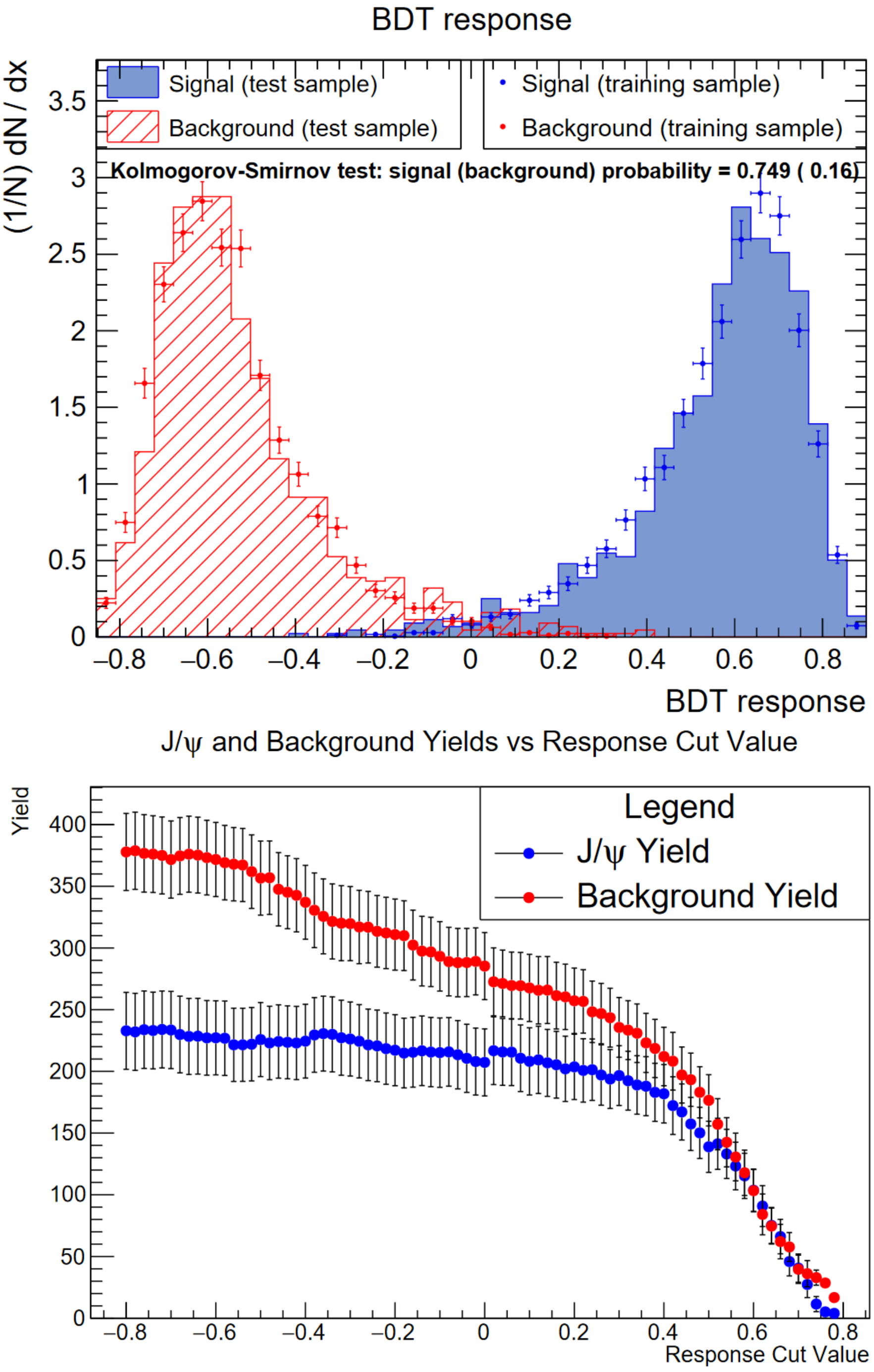


Figure 6.12: Top: The boosted decision tree (BDT) response when applied to the simulated training and testing sets, separated into the positron (blue) and pion (red) samples. Bottom: The J/ψ and background yield as a function of the lower limit cut on the response. The yields were evaluated on J/ψ photoproduction data taken with CLAS12 and not simulated data. Note that not all background will come from pions being misidentified as positrons.

applying a lower limit on the response of 0.3. The effect of the boosted decision tree is clearly observed when comparing these to the positron plots of Figure 6.9. The structure at low sampling fraction is mostly removed, and the plots are now more similar to the electron plots of Figure 6.9.

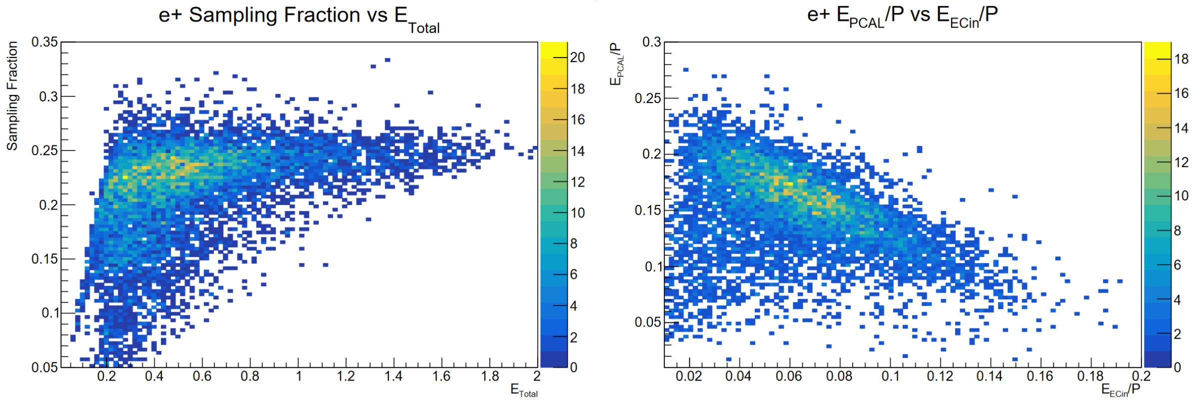


Figure 6.13: Right: The sampling fraction as a function of the sum of the energy deposited in the PCAL, ECIN and ECOUT. Left: The positron PCAL sampling fraction as a function of the ECIN sampling fraction. Both plots made with a cut on the positron boosted decision tree response of 0.3.

As seen when comparing Figure 6.11 to the bottom panel of Figure 6.12, tightening the sampling fraction cut for positrons achieves a similar ratio of the J/ψ to background yields as placing a cut at 0.3 on the response of the positron identification classifier. The later preserves more of the J/ψ yield than tightening the sampling fraction cut and so would be chosen due to the low J/ψ yields in the near-threshold region. Both positron identification procedures are however effective at reducing the amount of background and achieving a better J/ψ to background yield ratio and could be used almost interchangeably. This can further be useful in estimating the systematic uncertainty introduced by using one procedure rather than the other.

Once electrons and positrons are correctly identified, a few other analysis procedures are required. The same DC fiducial cuts as for muons and protons can be applied to remove tracks close to the edges of the drift chambers. Positrons and electrons will also often radiate a photon as they go through material between the target and detector. This is exemplified in the top left plot of Figure 6.14 where the blue dotted track corresponds to a photon reconstructed next to a positron track in magenta. Typically these photons will be found within a small polar angle difference, as shown by the peak close to zero in the difference between the electron and photon polar angle θ shown in the top right plot of Figure 6.14. As this can happen upstream of the drift chambers, the electron track will be bent by the magnetic field of the DC, whilst that is not the case for the photon. The azimuthal angle ϕ will therefore not necessarily be similar for the electron and the photon. These radiated photons can then be identified by a small θ difference with the electron, typically with $|d\theta| < 0.7$ rad. The momentum of the photon is then added back into that of the electron. As photons can be misidentified as neutrons, the correction is applied to all neutrals with $|d\theta| < 0.7$ rad relative to an electron or positron. More detail on the radiative corrections is found in Joseph Newton's Ph.D. Thesis [84]. Shown in the bottom plot of Figure 6.14 is the missing mass squared of an electron, positron and proton detected towards

J/ψ photoproduction. The missing mass of the reaction should here be the mass of the scattered electron, which is essentially zero at GeV energies. In Figure 6.14, both histograms peak at zero as expected. These were both made with the same events where a positron, electron and proton were detected. However, the blue histogram was made after correcting the electron and positron momentum for radiated photons. This has a narrower peak than the red histogram where the electron and positron did not have the radiative corrections. More events are also present in the tails of the missing mass squared distribution when the corrections are not applied.

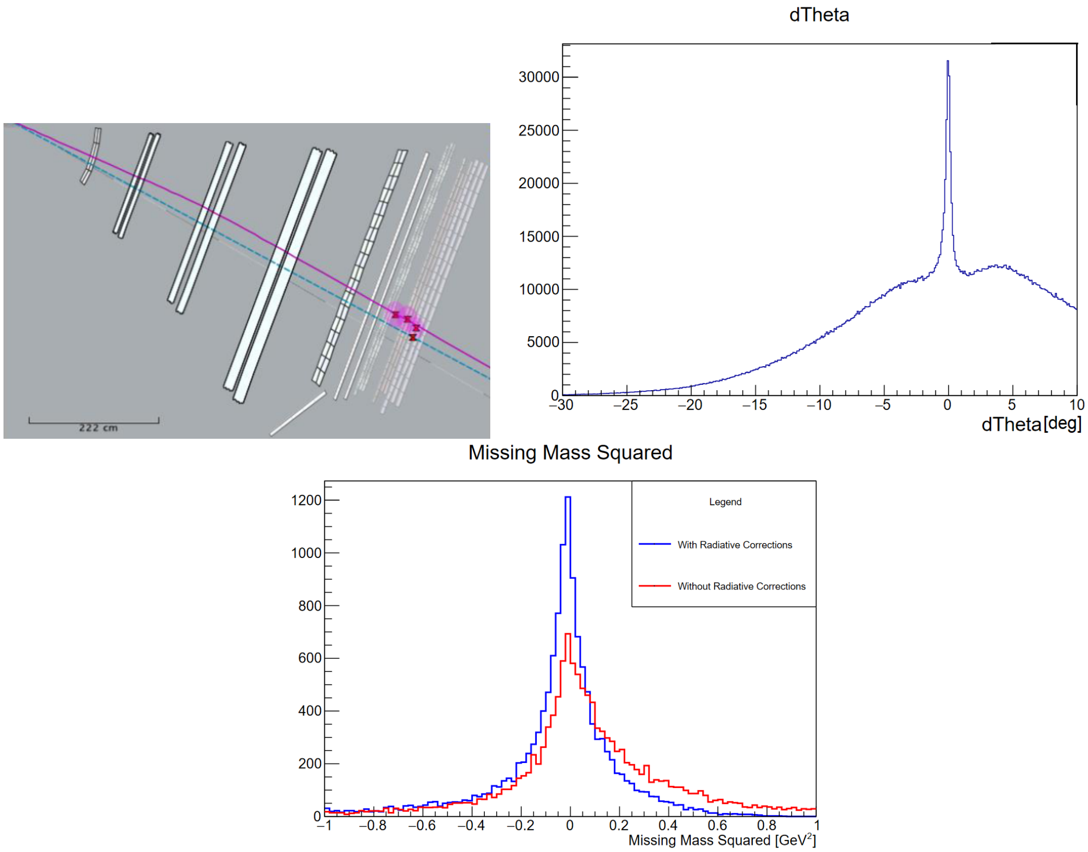


Figure 6.14: Top Left: A positron (magenta) produced in a simulation of a J/ψ event at CLAS12, shown in the CLAS12 Event Display (CED). A photon (blue) radiated by the positron was detected close to the positron. Top Right: The difference in θ between detected photons and electrons or positrons. A small difference of less then an a degree denotes photons radiated by an electron or positron. Bottom: The missing mass squared of a proton, positron and electron before (red) and after (blue) radiative corrections were applied to the positron and electron.

The radiative effects lead to two problems for the measurement of J/ψ photoproduction at CLAS12. Firstly good J/ψ events may be harder to fit in the invariant mass spectrum, as these may now be present in the tails of the J/ψ peak. The radiative effects will broaden the J/ψ peak and shift this away from the J/ψ mass, which again may make it harder to fit the J/ψ peak and therefore measure the J/ψ yield. These effects will be made explicit in Chapter 7 when comparing the invariant mass of two electrons and two muons that do not radiate photons.

Secondly, cuts are often applied to quantities such as the missing mass squared. When the resolution is made worse due to radiative effects, these cuts can start to be detrimental to the analysis, for example by cutting away good J/ψ events. The corrections detailed above allow to mitigate these effects. However, some photons will go undetected by CLAS12, and other errors in the reconstruction may also result in slight deviations in the measurement of electron or positron four-momentum.

6.4 Neutrons

Neutron detection and reconstruction is somewhat different to that of the particles described in the previous sections as neutrons are uncharged. This means that neutrons will not be detected in most of the FD's subsystems and they are only detected using a hit in the calorimeters. This then leads to several challenges for analyses requiring at least one neutron.

The first of these is due to the measurement of neutron momentum. As neutrons only interact with the calorimeters, the momentum is calculated from the reconstructed path length measured at the calorimeter, L_{cal}, the time of the hit at the calorimeter t_{cal} corrected for the start time, t_{st}, the mass of the neutron M_N and the speed of light c. First the velocity of neutrons β is calculated as:

$$\beta = \frac{L_{cal}}{c(t_{cal} - t_{st})} \tag{6.4}$$

The momentum of the neutron P is then calculated from the velocity as:

$$P = \frac{M_N \beta}{\sqrt{1-\beta^2}} \tag{6.5}$$

Past analyses with the predecessor of CLAS12, CLAS, showed that the path length of neutrons in the calorimeters is sometimes badly reconstructed [143], due to the thickness of the calorimeter and the unknown interaction point inside the calorimeter. The default average distance is taken as the front face of the calorimeters, whether this is the PCAL, ECIN or ECOUT. This is a good approximation for charged particles and photons that have a continuous interaction within the calorimeter. For charged particles this assumption is also not critical as the velocity is measured from the FTOF which has a better timing resolution than the calorimeters. For neutrons this approximation breaks down as neutrons do not have continuous interactions inside the calorimeter, and the hit will only be detected in a few strips of the calorimeters. Correcting this for neutrons is critical as the only measurement of momentum for neutrons comes from the calorimeters.

The $ep \to e'n\pi^+$ reaction is used to measure the difference between the average position of hits within the calorimeter and the approximation made that the hit occurs at the front face of the calorimeter in CLAS12 data. The neutron is detected and assigned a given path length based on the current approximation. No particle identification is used for the neutron meaning that the detected neutron candidates are simply all neutral particles in the event. Then the other decay products are used to calculate the neutron path length. By considering the missing energy E_X and missing momentum P_X calculated from the reaction $ep \to e'(n)\pi^+$ which does not take into account the neutron, the velocity of the missing neutron is calculated as:

$$\beta_X = \frac{P_X}{E_X} \tag{6.6}$$

The calculated path length is then given by:

$$L_X = c\beta_X(t_{cal} - t_{st}) \tag{6.7}$$

The difference between the reconstructed path length L_{cal} and the calculated path length L_X is used to correct the reconstructed path length. Slight errors in measuring the electron and pion momentum used to calculate β_X then lead to a worse resolution on the difference between path lengths. The mean of the difference is used as an average correction as this cannot be done on an event per event basis for every channel.

Several cuts are made to ensure that the calculated path length corresponds to the detected neutron. That is to say, that in the reaction $ep \to e'(n)\pi^+$ the missing particle is only a neutron and not a combination of additional particles. First, the polar angle of the missing particle is restricted between 5 and 35 degrees to ensure that the missing particle is going into the forward detector where the neutron is detected. Next, the cosine of the angle between the missing particle and the detected neutron is calculated. This peaks at 1, as shown in Figure 6.15, denoting that the missing particle has the same direction as the reconstructed particle. A lower limit on the cosine of the angle is placed at 0.995 to ensure that the missing particle corresponds to the detected neutron. Finally the missing mass of $ep \to e'(n)\pi^+$ is calculated. As shown in Figure 6.15, the missing mass presents a peak at the mass of the neutron, and cuts are made to restrict the mass range to between 0.85 and 1.05 GeV.

Figure 6.15 also compares the same variables calculated either on data taken with a hydrogen target or with a deuteron target. As shown the resolution is worse for a deuteron target. This is due to the fact that within a nuclear medium such as within the nucleus of a deuteron atom, the nucleons will have some inherent Fermi motion. This is not the case for hydrogen where the proton can be taken as free within the nucleus of a hydrogen atom. When calculating the missing mass of $ep \to e'(n)\pi^+$, the assumption is made that the proton inside the deuteron

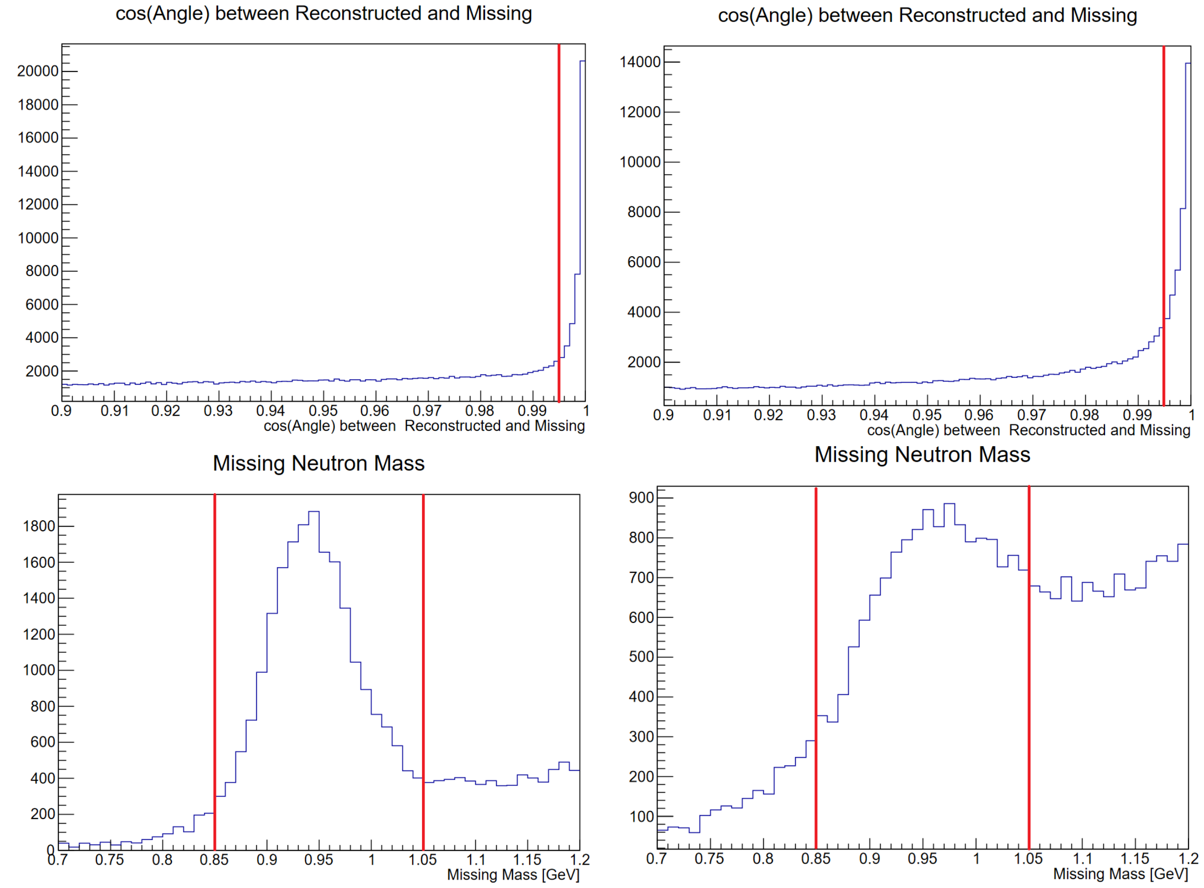


Figure 6.15: Top: The cosine of the angle between the missing particle and reconstructed neutron measured on a hydrogen (left) or a deuteron (right) target. Bottom: The missing mass of $ep \to e'(n)\pi^+$, with the cuts on the direction of the missing particle applied. This was measured on a hydrogen (left) or a deuteron (right) target. For all plots the red lines denote where the cuts have been placed.

target is stationary, whilst in fact it has some momentum which cannot be measured here. This assumption then leads to a worse resolution on the missing mass, or other quantities calculated for the missing particle, as the Fermi motion of the proton is not accounted for. Here this is relevant as it will lead to a worse resolution on the difference between reconstructed and calculated path lengths when measuring this on a deuteron target compared to a hydrogen target. Additional background for the deuteron target is due to the fact that more reactions can be produced that then become background for this analysis.

The difference between the calculated and reconstructed path length is calculated once the cuts detailed above are applied. The neutrons are required to have a hit in only one of the PCAL, ECIN, ECOUT, with the time used for the calculated path length taken from the same detector. The path length difference is then fitted with a gaussian. The extracted mean is then used as the correction to the reconstructed path length. Figure 6.16 shows the fitted path differences for three datasets, one taken on a hydrogen target with a beam energy of 10.2 GeV, the next taken on a deuteron target with a beam energy of 10.2 GeV and the final taken on a deuteron target with a beam energy of 10.6 GeV. As shown the gaussians fitted to the path length difference measured on data taken on a deuteron target have a larger sigma than those fitted to the difference measured on data taken with a hydrogen target. However there is good agreement between the datasets taken on a hydrogen and a deuteron target at the same beam energies.

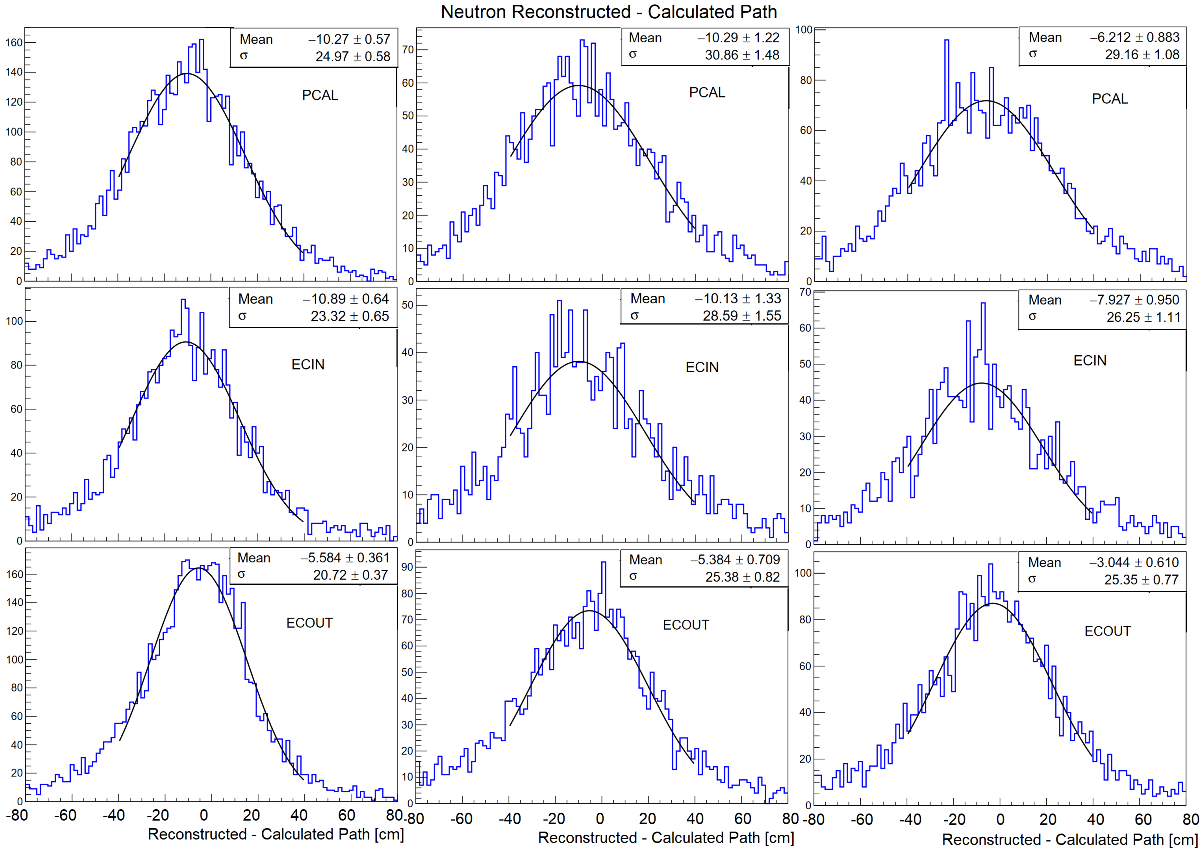


Figure 6.16: The difference between the calculated and reconstructed path length when the neutron is detected only in the PCAL (top row), ECIN (middle row) and ECOUT (bottom row). The difference in path length is measured on three datasets, first with a dataset taken with a hydrogen target and beam energy of 10.2 GeV (left column), next with a dataset taken with a deuteron target and beam energy of 10.2 GeV (middle column) and finally with a dataset taken with a deuteron target and beam energy of 10.6 GeV (right column).

The fact that the difference in the path length is larger when comparing the datasets taken with a beam energy of 10.2 GeV to the one taken with a beam energy of 10.6 GeV indicates a momentum dependence of the path length corrections. The average neutron momentum at 10.2 GeV is different than that at 10.6 GeV, which explains the difference in the path length corrections. As the path corrections aim to correct the momentum of neutrons it is hard to establish a momentum dependent path length correction. One solution would be to create a probability matrix, where the probability that the reconstructed momentum corresponds to a calculated momentum is created for bins of reconstructed and calculated momentum. The correction would then be calculated in the same bins of calculated momentum. When applying the correction, the correction from the bin in calculated momentum which has the highest probability of corresponding to the reconstructed momentum would be used. This however has not been established due to the fact that changes in the reconstruction of the neutron path length based on this work are going to be implemented in a new release of CLAS12 data which will affect the correction to the path length. Further development of the path length corrections will be done once this new release is available.

Similar corrections can be established from simulated data. In this case, the usual J/ψ simulations can be used instead of relying on the $ep \to e'n\pi^+$ channel. Instead of calculating a missing neutron β as before, the true generated β value for the neutron is used to calculate the path length. The detected neutron is matched to the true neutron by taking the particle with four momentum closest to that of the true neutron. The path length is then calculated as before. Figure 6.17 then shows measured path length differences in simulation. The smaller difference in correction between the datasets simulated with a beam energy of 10.2 GeV or 10.6 GeV compared to that extracted from the $ep \to e'n\pi^+$ channel in CLAS12 data indicates that there is a smaller difference in average momentum in J/ψ photoproduction than in the $ep \to e'n\pi^+$ channel. The much larger corrections established from the simulated data compared to that established from CLAS12 data is due to start time calibrations that are applied in CLAS12 data but not in simulated data, as these calibrations are not needed in simulation where the start time is exactly known. As the calculated path length from Equation 6.7 depends on the start time, the path length correction for CLAS12 data also corrects for the start time calibrations.

As mentioned before, the corrections are then taken as the mean of the gaussian fit for each calorimeter. An average of the corrections on the 10.2 GeV and 10.6 GeV datasets are used. The corrections were established for neutrons with a hit in only one of the calorimeters, however neutrons can interact with more than one of the calorimeters. In that case, the path length, time and corrections are taken from the calorimeter closest to the target which goes in the order of PCAL then ECIN then ECOUT. The corrected β is then calculated as:

$$\beta = \frac{L_{cal} - L_{cor}}{c(t_{cal} - t_{st})} \tag{6.8}$$

for L_{cor} the correction measured in Figures 6.16 and 6.17. The effect of the corrections on simulated data is shown in Figure 6.18. As shown, the corrected β is much closer to the true generated β than the uncorrected β. Figure 6.18 also shows the effect of the path length corrections on the missing mass squared of the $ep \to (e')e^+e^-n$ channel used for the J/ψ analysis. The path length corrections increase the peak close to zero and reduce some of the tails at negative missing mass squared. The effect is not drastic due to the small corrections extracted from Figure 6.16. However, given the large corrections needed in simulated data, applying these corrections to both

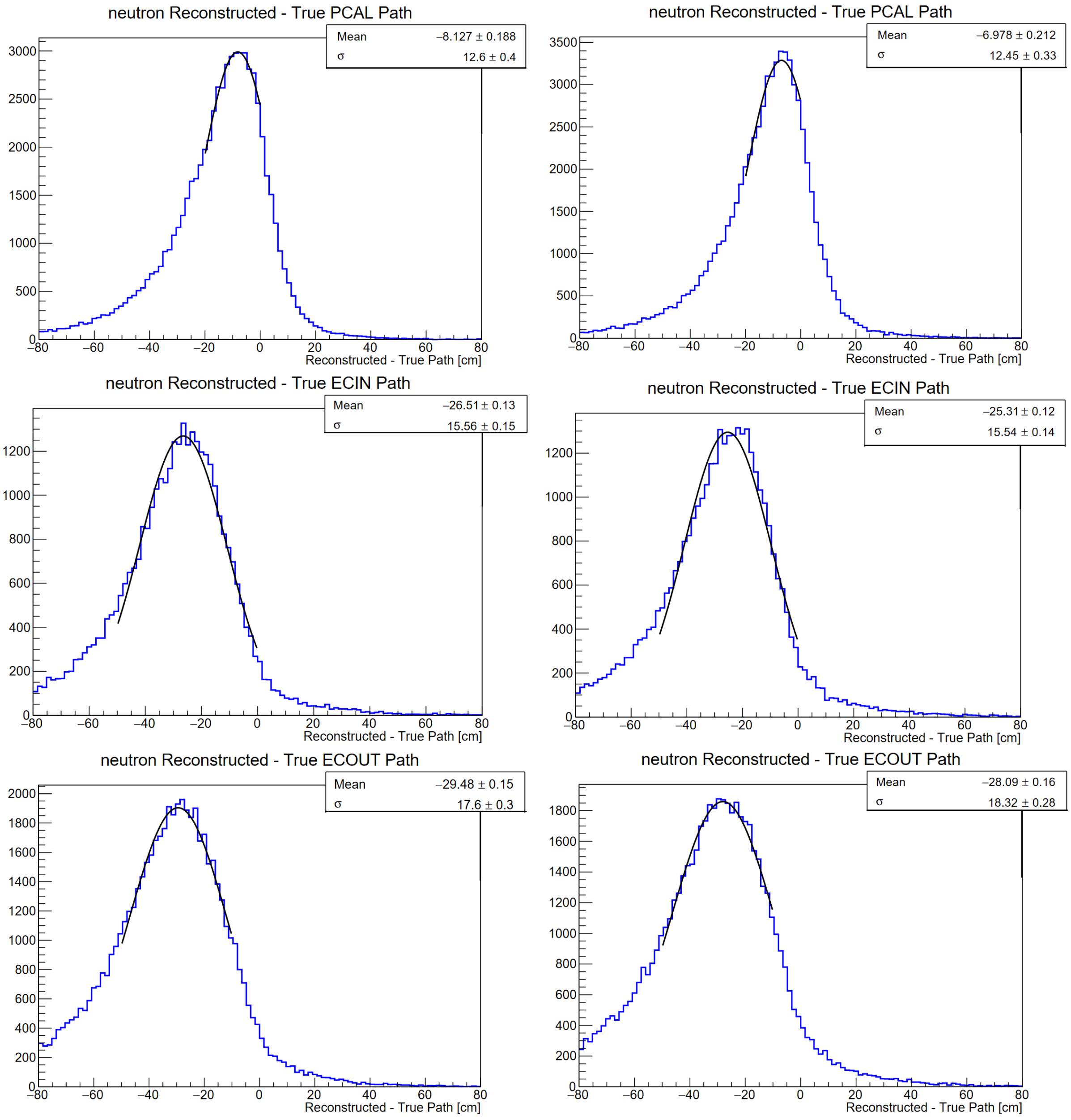


Figure 6.17: The difference between the true and reconstructed path length when the neutron is detected only in the PCAL (top row), ECIN (middle row) and ECOUT (bottom row). The difference in path length is measured on two simulated datasets, first with a dataset simulated with a deuteron target and beam energy of 10.2 GeV (left column) and next with a dataset simulated with a deuteron target and beam energy of 10.6 GeV (right column).

simulated and CLAS12 data ensures better agreement between the two.

The next challenge of neutron detection at CLAS12 is made explicit in Figure 6.19. This shows a single sector of the CLAS12 detector in the CLAS12 Event Display (CED). In principle this sector should only contain one neutron, as the event was produced in a simulation of J/ψ photoproduction on the neutron and the simulation only contains one neutron. However, the reconstruction algorithm identified several neutron candidates, denoted in green tracks, and a

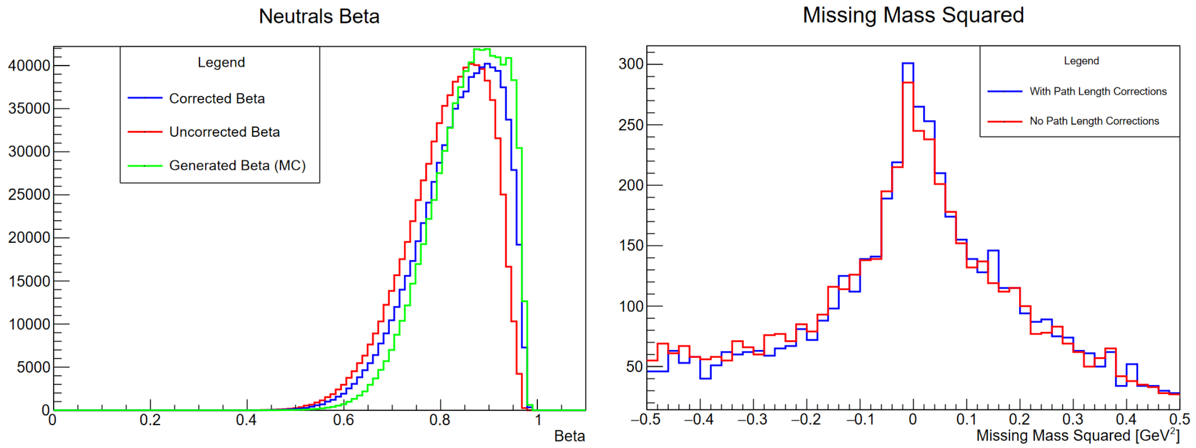


Figure 6.18: Left: Neutron velocity with (blue) and without (red) the path corrections applied and compared to the true β of the simulated neutron (green). Right: The missing mass squared of $ep \to (e')e^+e^-n$ with (blue) and without (red) the path length corrections.

photon candidate, denoted in the blue track. To make the problem explicit, multiple neutral candidates occur for many (although not all) neutrons passing through CLAS12.

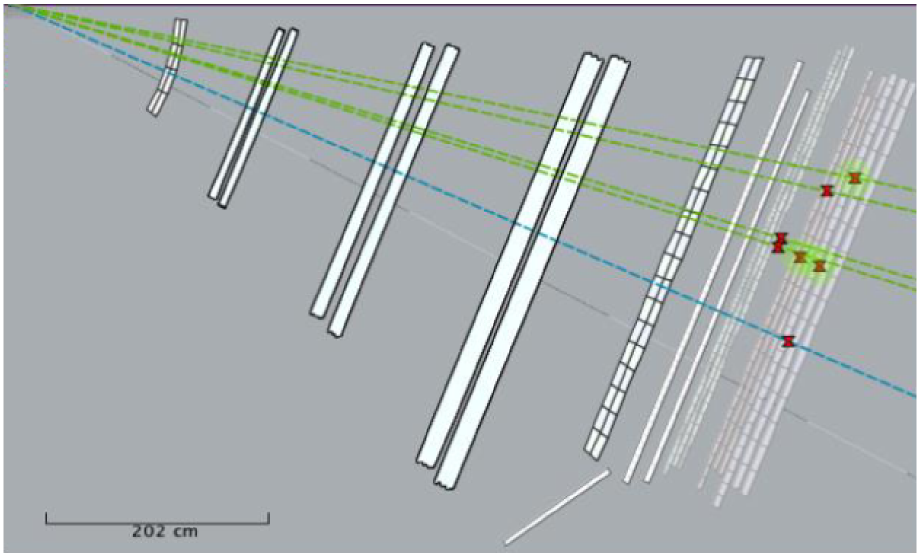


Figure 6.19: A sector of the CLAS12 detector containing multiple neutron tracks (green) and a photon track (blue), shown in the CLAS12 Event Display (CED). The event was produced in a simulation of J/ψ photoproduction on the neutron and should only contain one single neutron.

This problem can be analysed in simulation by considering the neutron with four-momentum closest to the true generated neutron as the primary or truthmatched neutron. In CLAS12 data, the $ep \to e'n\pi^+$ channel is used to identify the primary or truthmatched neutron as the one closest to the missing neutron calculated from $ep \to e'(n)\pi^+$. Most secondary neutrons are within the same sector as the primary neutron. A first hint as to what might be happening is shown in Figure 6.20. This shows the distance between the hits in a calorimeter of the truthmatched neutron and any secondary neutron in the same event and firing the same calorimeter. The neutrons are required to only interact with one of the PCAL (blue), ECIN (red) or ECOUT (green). The spikes in the diagram correspond to a distance of one, two or three strips, with the strips 4.5 cm wide in the PCAL and 9 cm wide in the ECIN and ECOUT. This is probably due to the peak and cluster finding algorithms of the calorimeters that erroneously create multiple clusters associated with only one neutron. As no tracking information is available to correlate the clusters to a track, these multiple clusters are then all associated with different neutrons.

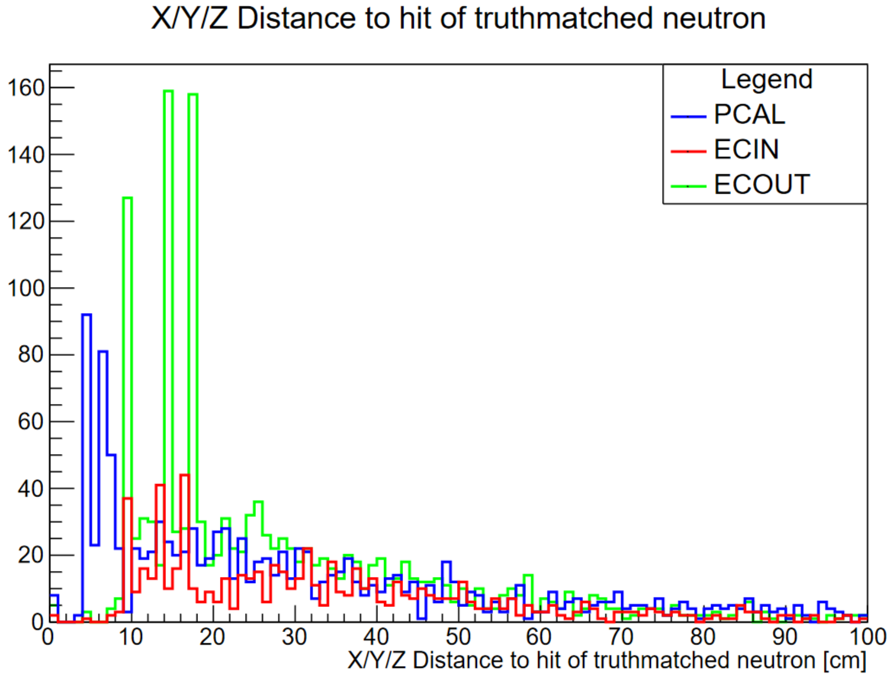


Figure 6.20: The distance between the calorimeter hits of the truthmatched neutron and that of any secondary neutron in the same event and firing the same calorimeter. The neutrons are required to only interact with one of the PCAL (blue), ECIN (red) or ECOUT (green).

A second hint is given in Figure 6.21. This shows two different scenarios. In the top one, a low momentum neutron interacts first with the PCAL, before interacting with the ECAL, knocking out a proton which then creates several clusters. The cluster matching algorithm in principle should assign only one track to all clusters, as it does in this example denoted by the thin black line connecting the clusters, but may fail and create several separate neutrals. In the bottom scenario, a neutron interacts with the FTOF, knocking out a proton and producing a neutral pion. The neutral pion then decays into two photons that create two separate sets of clusters identified as two different tracks. The FTOF information is unfortunately not preserved for clusters that are not associated to charged tracks in the DC. However if it was, then up to four neutrals could have been created from the clusters produced here.

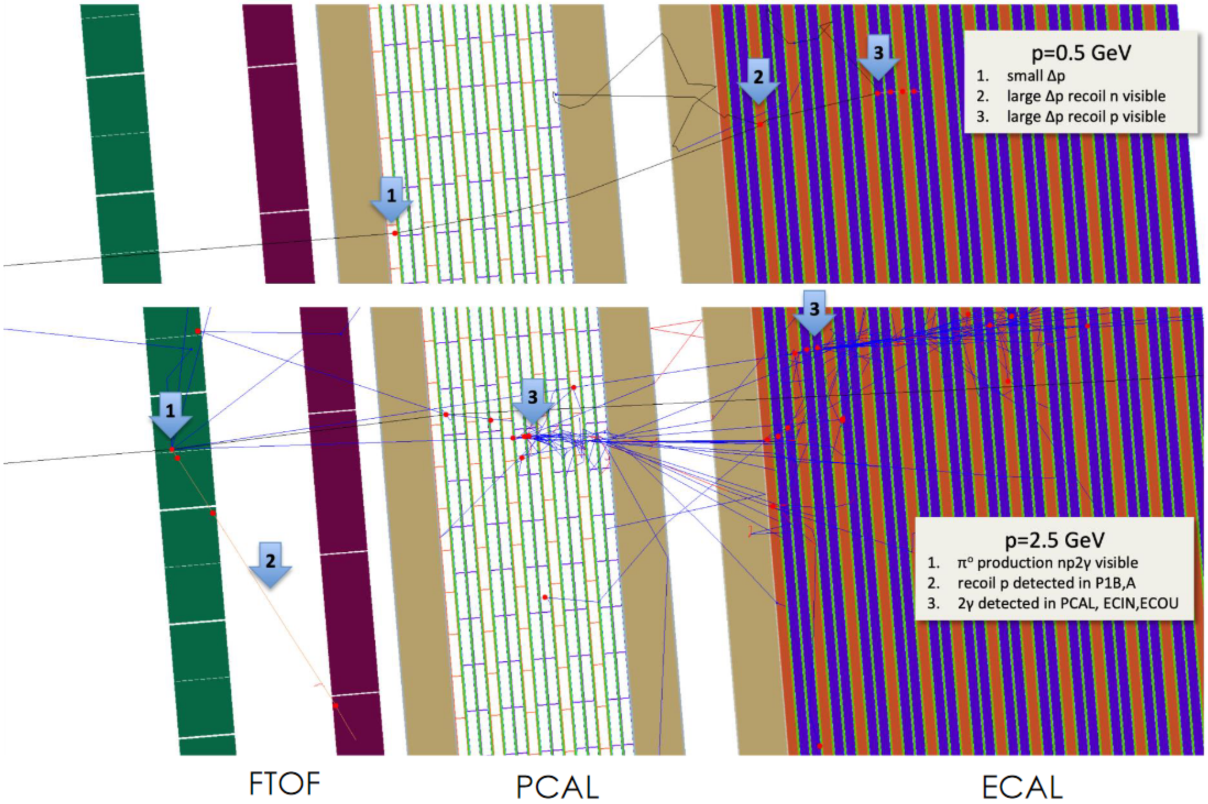


Figure 6.21: Examples of real physical processes that lead to multiple clusters produced by a single neutron. Detailed description of the physical processes in the legend.

As such, whether due to real physical processes or errors in the reconstruction, a single neutron track can lead to multiple reconstructed neutrons at CLAS12. This effect is seen in both simulation and real data. In channels such as $ep \to e'n\pi^+$ where all final state particles are detected, neutron candidates can be matched to the missing neutron calculation as explained before. However this is not possible for J/ψ photoproduction as the scattered electron is not detected, as explained in Chapter 2. Two other ways to select the primary or best neutron can be established. First by considering a χ^2 as:

$$\chi^2 = \Sigma_{variables} \frac{(measured - expected)^2}{expected} \tag{6.9}$$

which is calculated by taking the difference between the measured value of a variable and what would be expected, and summing this over several different variables. For example, for J/ψ photoproduction, the missing mass of the event should be the mass of an electron. The χ^2 can then be calculated for each neutron in an event by taking the measured value of the missing mass and comparing it to the mass of an electron. The neutron that minimises this χ^2 value is then taken as the primary neutron.

Another possibility is to take the earliest neutron. This selection is made within loose cuts to ensure only reasonable candidates are considered, for example removing noise in the calorimeters that was reconstructed as a neutral particle. Figure 6.22 shows the time difference between the truthmatched or best neutron and secondary neutrons, when these are both detected in only one the PCAL/ECIN/ECOUT. As shown, the time difference is mostly negative, indicating that the truthmatched neutron is typically the earliest. When the truthmatched and secondary neutron are in different calorimeters, the same rule applies. Typically this also means that the neutron with a hit closest to the target is selected, although in the case of a secondary neutron travelling towards the target, its time would still be greater than that of the primary neutron.

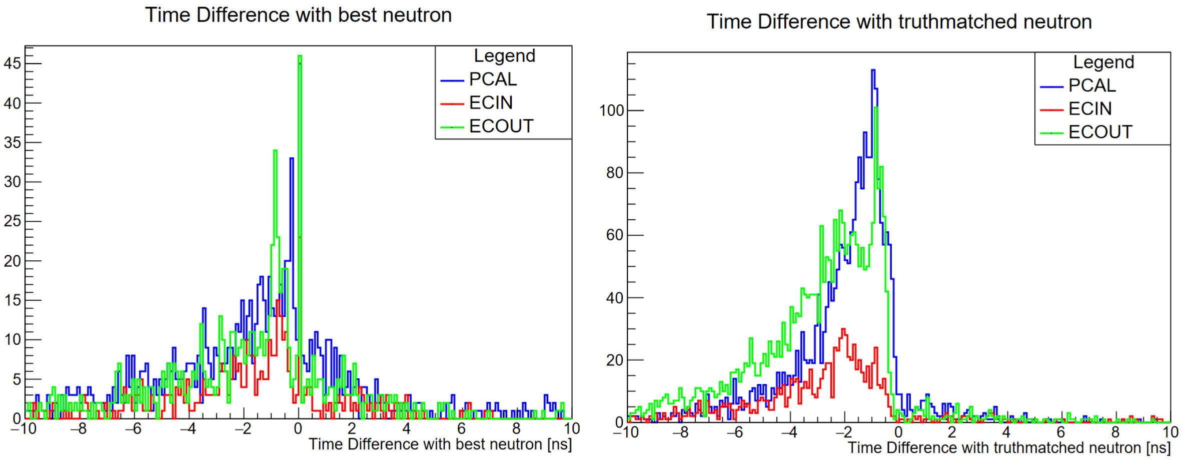


Figure 6.22: The difference in time between the truthmatched or best neutron and secondary neutrons for data taken with CLAS12 (left) and simulated data (right).

The effect on the neutron β is shown in Figure 6.23. The time ranking method of selecting the best neutron results in a slightly higher beta overall, which is logical given that the earliest neutral will be the fastest. On the other hand, the χ^2 ranking is biased towards the neutrons that create the best measurements of the variables used in the ranking. As such preference was given to the time ranking method as it seemed less likely to introduce bias into the analysis. Also shown in Figure 6.23 is the histogram made when considering all neutrons in an event. These contribute to roughly 30% more entries, denoting a sizeable effect that warrants the effort described above in suppressing these secondary neutrons.

Finally, the last consideration for neutrons in the analyses of J/ψ photoproduction at CLAS12 is the need to distinguish between neutrons and photons. Typically at CLAS12 neutrals are identified as either photons or neutrons based on their velocity. Neutrals with β close to one are identified as photons, and all others are identified as neutrons. However, as shown in Figure 6.23 and Figure 6.18, neutron candidates in J/ψ photoproduction data have β very close to one. As such cutting on β to remove photon contamination is unfeasible.

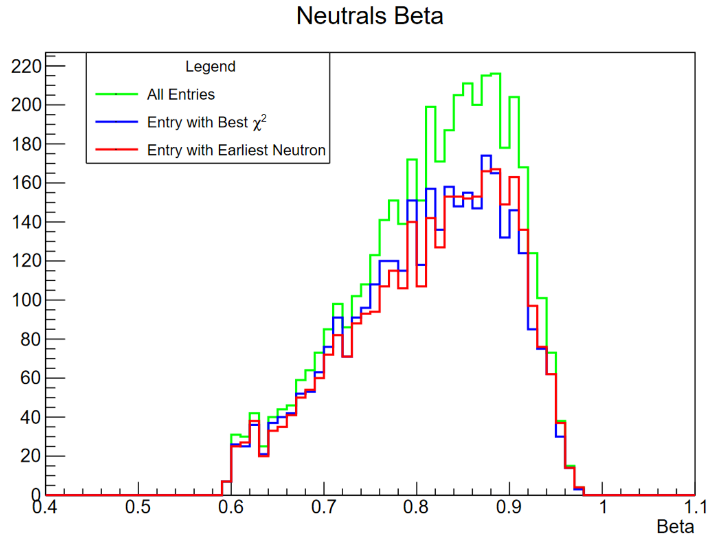


Figure 6.23: The neutron β measured on J/ψ photoproduction data at CLAS12, when selecting the neutron candidate that maximises the χ^2 ranking (blue), with the earliest time (red) or when taking all neutrons in an event (green).

Another solution would be to train a machine learning algorithm as in the previous sections to identify neutrons and remove photons. However, some qualitative arguments seem to indicate that there is not a lot of photon contamination in the current J/ψ photoproduction data. Figure 6.24 first shows the energy deposited in the PCAL by simulated neutrons, simulated photons and all neutrals considered in J/ψ photoproduction data at CLAS12. Photons should have much higher energy deposition than neutrons as they produce showers via pair production and Compton scattering in the calorimeters. This is shown for the simulated photons in Figure 6.24 whilst the neutrals in CLAS12 data do not have a high energy deposition. Figure 6.24 also shows the number of calorimeter subsystems that were hit by simulated photons, neutrons and neutrals in CLAS12 data. These subsystems refer to the PCAL, ECIN, ECOUT. As shown, photons are much more likely to interact with multiple calorimeter subsystems than neutrons. Again, the neutrals in CLAS12 data seem to behave more like the simulated neutrons. Overall it was decided that no particle identification other than charge would be required for neutron candidates, based on the above observations. From now on, neutrons can be thought of as any neutral particle to which the above mentioned correction and selection criteria was applied.

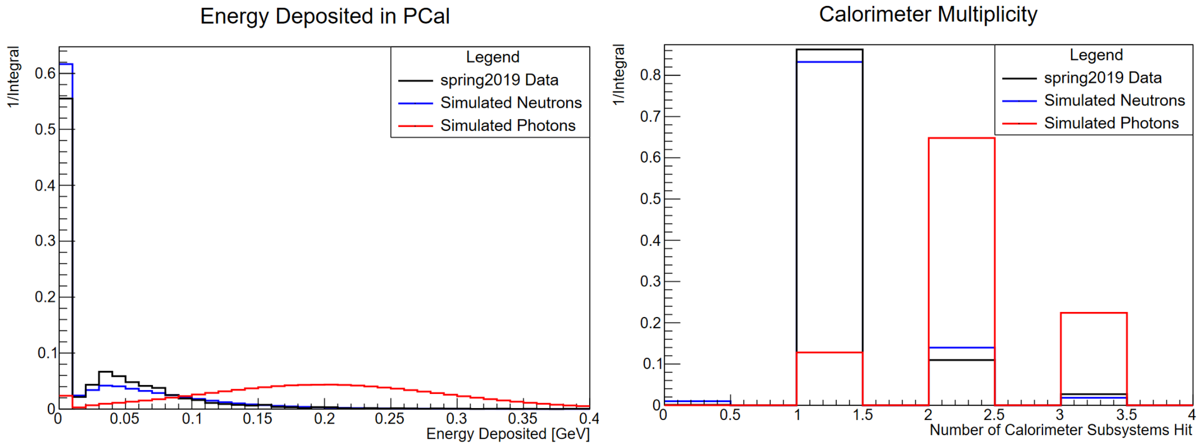


Figure 6.24: Left: The energy deposition in the PCAL of simulated neutrons (blue), simulated photons (red) and neutrals in J/ψ photoproduction data at CLAS12 (black). Right: The number of calorimeter subsystems hit for simulated neutrons (blue), simulated photons (red) and neutrals in J/ψ photoproduction data at CLAS12 (black). The subsystems refer to the PCAL, ECIN, ECOUT. For both Left and Right the histograms are normalised by one over their integral, such that the sum of each bin gives one.

Chapter 7

Reaction Identification

Once final state particles have been detected and identified, and their four momentum has been corrected, the next step is to identify the reaction that produced them. In the case of J/ψ photoproduction at CLAS12, the final state particles are composed of a lepton pair and a recoil nucleon, with the scattered electron going undetected as mentioned in Chapter 2. The first section of this chapter will describe how cuts on the missing mass squared and Q^2, as defined by equations 2.5 and 2.1 respectively, can be used to identify di-lepton photoproduction events. These include J/ψ photoproduction but also other channels where a di-lepton pair is produced by a quasi-real virtual photon interacting with the target, such as, for example, Bethe Heitler or Timelike Compton Scattering [141,142]. The next section will detail how machine learning can be used to select J/ψ photoproduction events and remove these other channels. The third section of this chapter will focus on the di-lepton invariant mass spectrum. The fourth section will then demonstrate that the analysis so far is able to select only exclusive J/ψ photoproduction events. The final section will then show a very preliminary measurement of the unpolarised J/ψ SDMEs.

7.1 Cutting on the Missing Mass Squared and Q^2

Defined in equation 2.5, the missing mass squared of the channel $eN \rightarrow (e')l^+l^-N$ should be the mass squared of the scattered electron that goes undetected, as indicated by the parentheses. At GeV energies the mass squared of an electron is essentially zero. Here l^+l^- denotes the lepton pair, with N denoting the nucleon on which the di-lepton pair was produced. An example of the missing mass squared for the $ep \rightarrow (e')\mu^+\mu^-p$ is shown in Figure 7.1. In this case the reaction requires two final state muons and a proton, as defined in Chapter 6. This channel is produced on a proton contained in a hydrogen target. From now on p and p_{bound} are used to refer to the free proton in a hydrogen target and the bound proton in a deuteron target. The missing mass squared in Figure 7.1 shows a peak at zero as expected. The missing mass squared can then be used to remove events that would have additional missing particles X in for example

$ep \to (e')\mu^+\mu^- pX$. These events are then removed by applying cuts around the peak at zero.

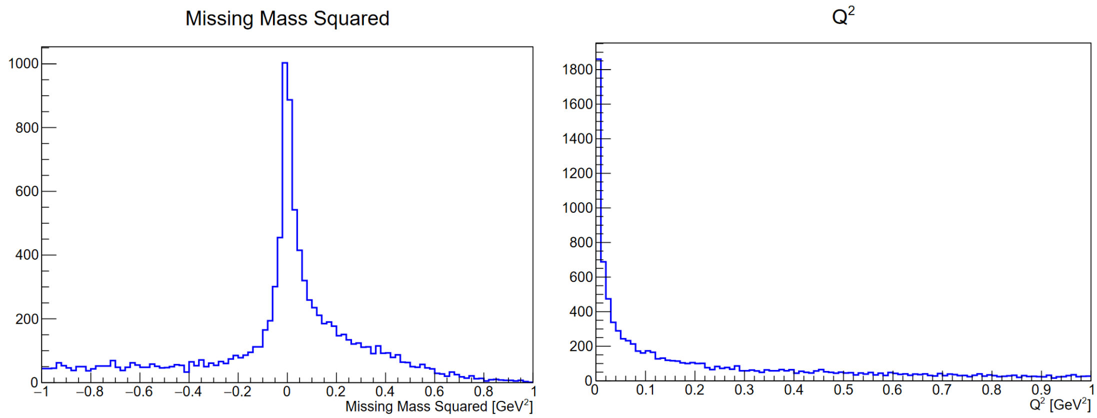


Figure 7.1: Left: The missing mass squared of $ep \to (e')\mu^+\mu^- p$. Right: Q^2 of $ep \to (e')\mu^+\mu^- p$ events. Both plots peak at zero as expected.

Figure 7.1 also shows Q^2 for the same $ep \to (e')\mu^+\mu^- p$ events. Q^2 corresponds to the mass of the virtual photon mediating the interaction, as shown in Figure 2.1. For quasi-real photoproduction, the mass of the photon and therefore Q^2 will be almost zero. Cutting on Q^2 removes non photoproduction events and background. Another set of variables that can be used to identify quasi-real photoproduction events are based on the direction of the missing scattered electron. The momentum fraction in the transverse X and Y components of the scattered electron momentum can be calculated by considering the missing momentum of the reaction $ep \to (e')\mu^+\mu^- p$. The total transverse momentum fraction is then given by [144]:

$$\frac{P_t^{miss}}{P^{miss}} = \sqrt{\left(\frac{P_X^{miss}}{P^{miss}}\right)^2 + \left(\frac{P_Y^{miss}}{P^{miss}}\right)^2} \tag{7.1}$$

Quasi-real photoproduction events will have a very forward going scattered electron, essentially meaning that it will have a very low transverse momentum fraction. Figure 7.2 shows the transverse momentum fraction of the missing scattered electron as a function of Q^2. As shown events at low Q^2 also have a low transverse momentum fraction. Both variables are correlated and so typically a cut is placed on either Q^2 or the missing transverse momentum. Note that $\frac{P_t^{miss}}{P^{miss}}$ is written as $\frac{P_t}{P}$ for brevity. Figure 7.2 also shows a plot of the X versus Y components of the missing transverse momentum fraction. These are uncorrelated and can essentially be used as two independent cuts to select quasi-real photoproduction events.

Figure 7.3 shows the J/ψ and background yields as a function of the cut width from 0 on the absolute value of the missing mass squared (MM^2) and Q^2. This is shown for the channels where J/ψ is produced on a hydrogen target and decays to a di-muon pair, and when J/ψ is produced on either the bound proton or neutron in a deuteron target and decays to a di-electron pair. Data

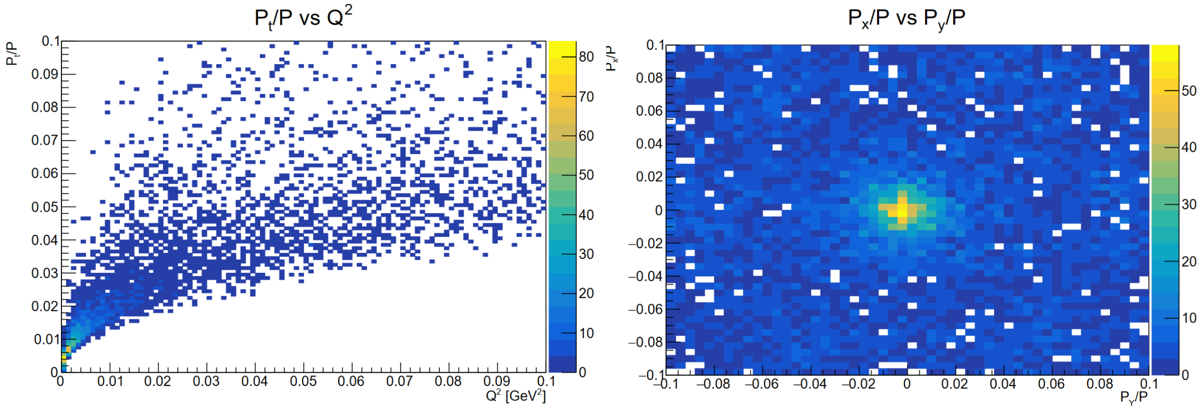


Figure 7.2: Left: The missing transverse momentum fraction as a function of Q^2 for $ep \rightarrow (e')\mu^+\mu^- p$ events. Right: The x and y components of the missing transverse momentum fraction plotted against one another.

taken on a deuteron target typically has a larger background due to additional available channels that contribute to the background. The missing mass squared and Q^2 also typically have a better resolution on a hydrogen target, as mentioned in Chapter 6 due to the Fermi motion of the constituents of the deuteron. Finally, neutrons typically have a lower detection efficiency than protons explaining why the channel produced on the bound neutron has a lower J/ψ yield. Figure 7.3 can then inform the decision of where to place a cut on the missing mass squared or Q^2. Typically for these analyses, these cuts are placed around 0.3 GeV^2, although the cut can be varied to study the systematic effect this has on measured quantities such as the cross section.

The resolution of the missing mass squared, Q^2, $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ was measured as a function of the quasi-real photon energy E_γ, by fitting a gaussian distribution to the peak of these variables binned in E_γ. A parametrisation is then extracted from the σ of the gaussian as a function of E_γ. As shown in Figure 7.4, there is a strong dependency of the resolution of these variables on E_γ. This holds true across different channels, measured on a proton, bound proton or bound neutron, and where a di-electron, di-pion or di-muon pair were produced. Note that the resolution of Q^2 was not extracted for the channels produced on a free proton; at the time when these analyses were made the preference was given to cuts on $\frac{P_X}{P}$ and $\frac{P_Y}{P}$. In practice cutting on either Q^2 or $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ will lead to similar results.

As mentioned in Chapter 6, path length corrections are applied to correct the momentum reconstruction of the neutron. These fail to capture the momentum dependence of the path length correction, as establishing a momentum dependence of the path length correction which aims to correct momentum quickly becomes a circular problem. A consequence of this is that the path length corrections are devised for the mean momentum of the neutron sample used to calculate the correction. Figure 7.5 shows how the mean of the missing mass squared peak is shifted away from zero at lower quasi-real photon energy, due to the path length correction not

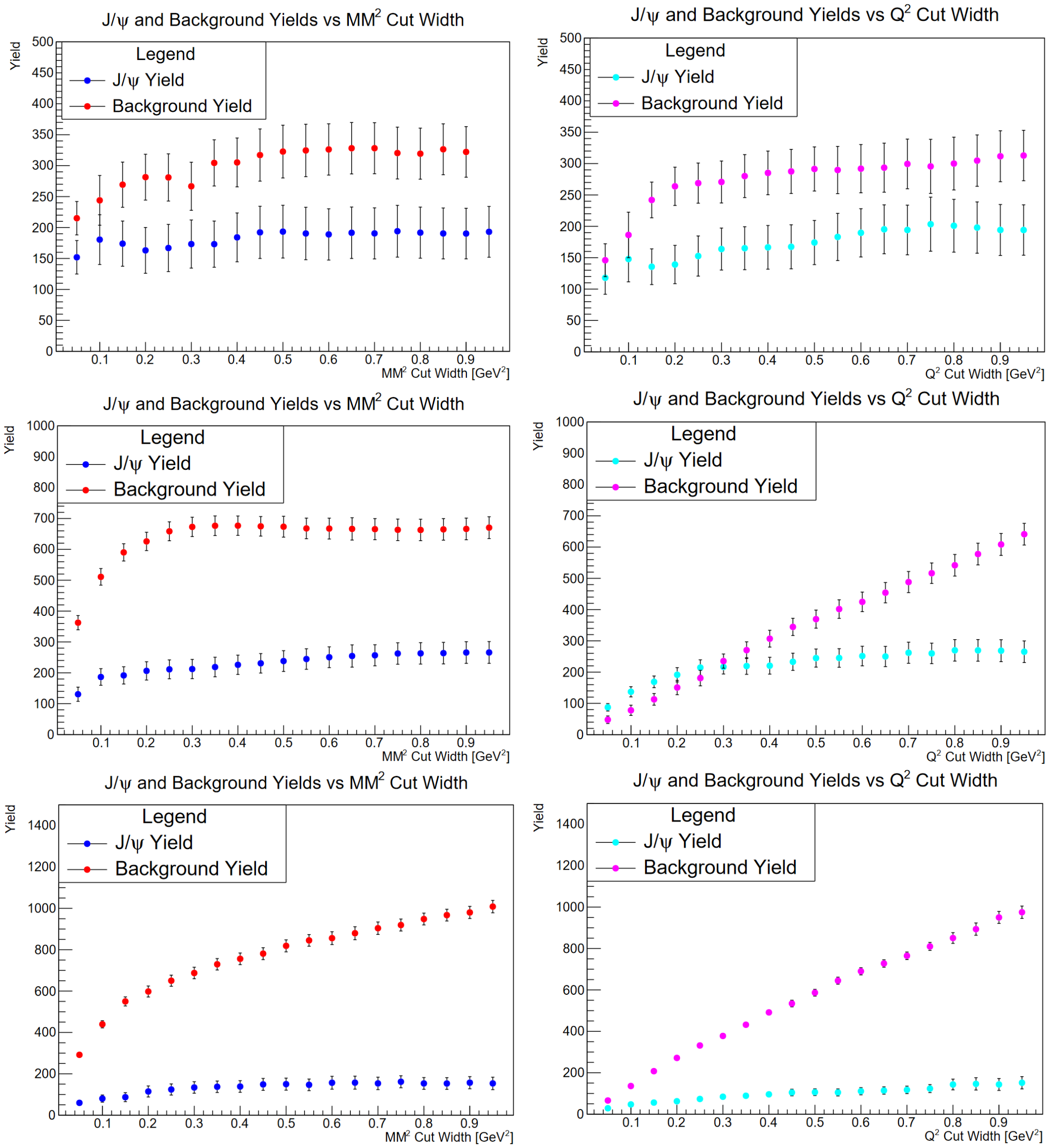


Figure 7.3: The J/ψ and background yields as a function of the cut width from 0 on the absolute value of the missing mass squared (Left) and Q^2 (Right), for the channels $ep \rightarrow (e')\mu^+\mu^-p$ (top row), $ep_{bound} \rightarrow (e')e^+e^-p$ (middle row) and $en_{bound} \rightarrow (e')e^+e^-n$ (bottom row).

being suited for the mean neutron momentum at lower E_γ. This shift in the mean position of the peak is fitted so as to apply a parametrisation. The resulting cut on the missing mass squared for the $en_{bound} \rightarrow (e')e^+e^-n$ channel is then corrected for this shift as a function of E_γ.

Figure 7.4 shows the parametrisations for the resolution of the missing mass squared, Q^2, $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ in simulated data. Figure 7.6 shows the ratio between experimental and simulated data for the missing mass squared resolution . This shows an almost flat dependency up until around 10 GeV where the ratio diverges, due to the parametrisations of the resolution dropping close to zero. Unfortunately it is hard to extract the resolution for Q^2, $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ in CLAS12 data

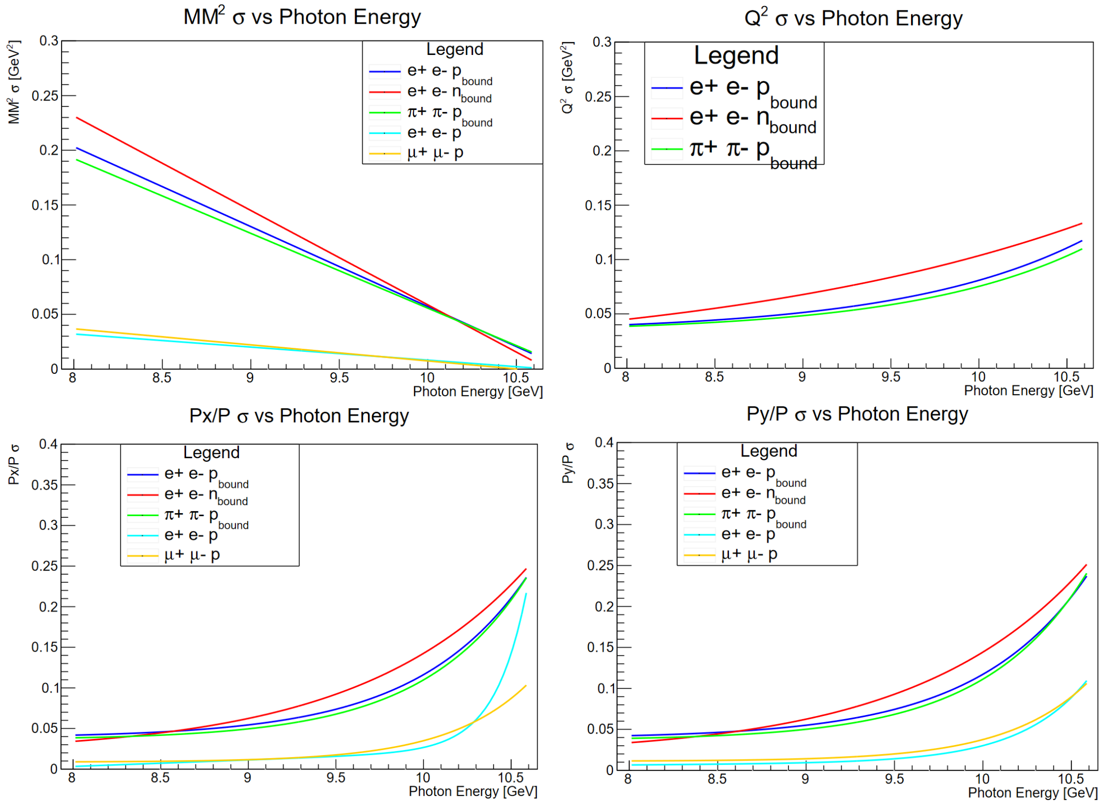


Figure 7.4: The resolution of the missing mass squared (top left), Q^2 (top right), $\frac{P_X}{P}$ (bottom left) and $\frac{P_Y}{P}$ (bottom right) as a function of the energy of the quasi-real photon involved in the interaction. These plots were produced on simulated data.

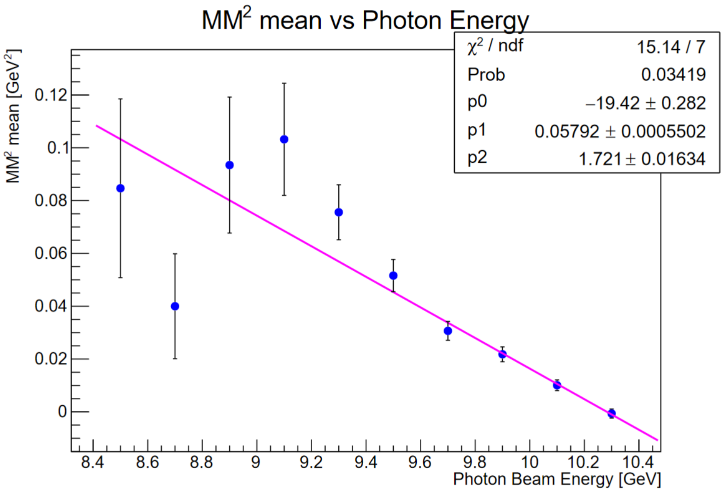


Figure 7.5: The mean of the missing mass squared peak for the $en_{bound} \rightarrow (e')e^+e^-n$ channel, when binned in E_γ. At lower E_γ the mean is shifted away from zero. This is corrected by a parametrisation for the shift in the peak mean.

taken on a deuteron target due to a large electroproduction background. Instead the analyses can use the parametrisations established on simulated data for Q^2, with the expectation of having a similarly larger resolution in CLAS12 data compared to simulated data as shown for the missing mass squared in Figure 7.6.

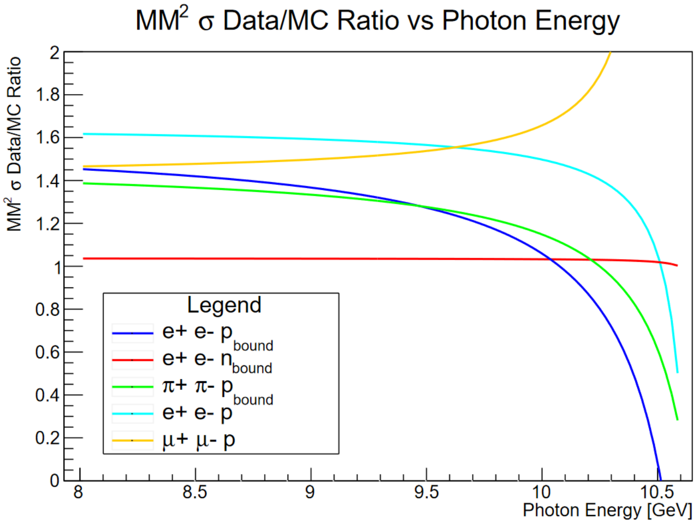


Figure 7.6: The ratio of the resolution of the missing mass squared in CLAS12 data over simulated data.

The parametrisations of Figure 7.4 can then be used to cut on the missing mass squared, Q^2, $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ to select di-lepton photoproduction events as detailed previously. By comparing Figures 7.3 and 7.7 it was established that the parametrised cuts on the missing mass squared allow to preserve more J/ψ yield whilst decreasing the ratio of J/ψ yield to background. The cuts on Q^2 for the deuteron target channels do not seem as effective, this might be due to the parametrisation established from simulated data. Instead a straight cut on Q^2 is preferred to the parametrisations. Finally, for the di-muon channel, the parametrised cut on $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ preserve more J/ψ yield and are preferred to straight cuts on Q^2.

7.2 Machine Learning Based Reaction Identification

The previous section has detailed how cuts on quantities like the missing mass or Q^2 can be used to select di-lepton photoproduction events. Extra care was placed in optimising these cuts due to a low J/ψ yield that needs to be preserved in order to achieve a statistical precision as good as possible when extracting quantities such as the cross section. Similarly, these cuts need to be able to remove the large backgrounds present in some of the different channels. None-negligible resolutions on the missing mass or Q^2 that vary as a function of the kinematics of the event further complicate this task. One can instead think of using machine learning for a classification task such as this one.

The machine learning approach will aim to specifically select J/ψ photoproduction events, instead of selecting a broader set of di-lepton photoproduction events of which J/ψ photopro-

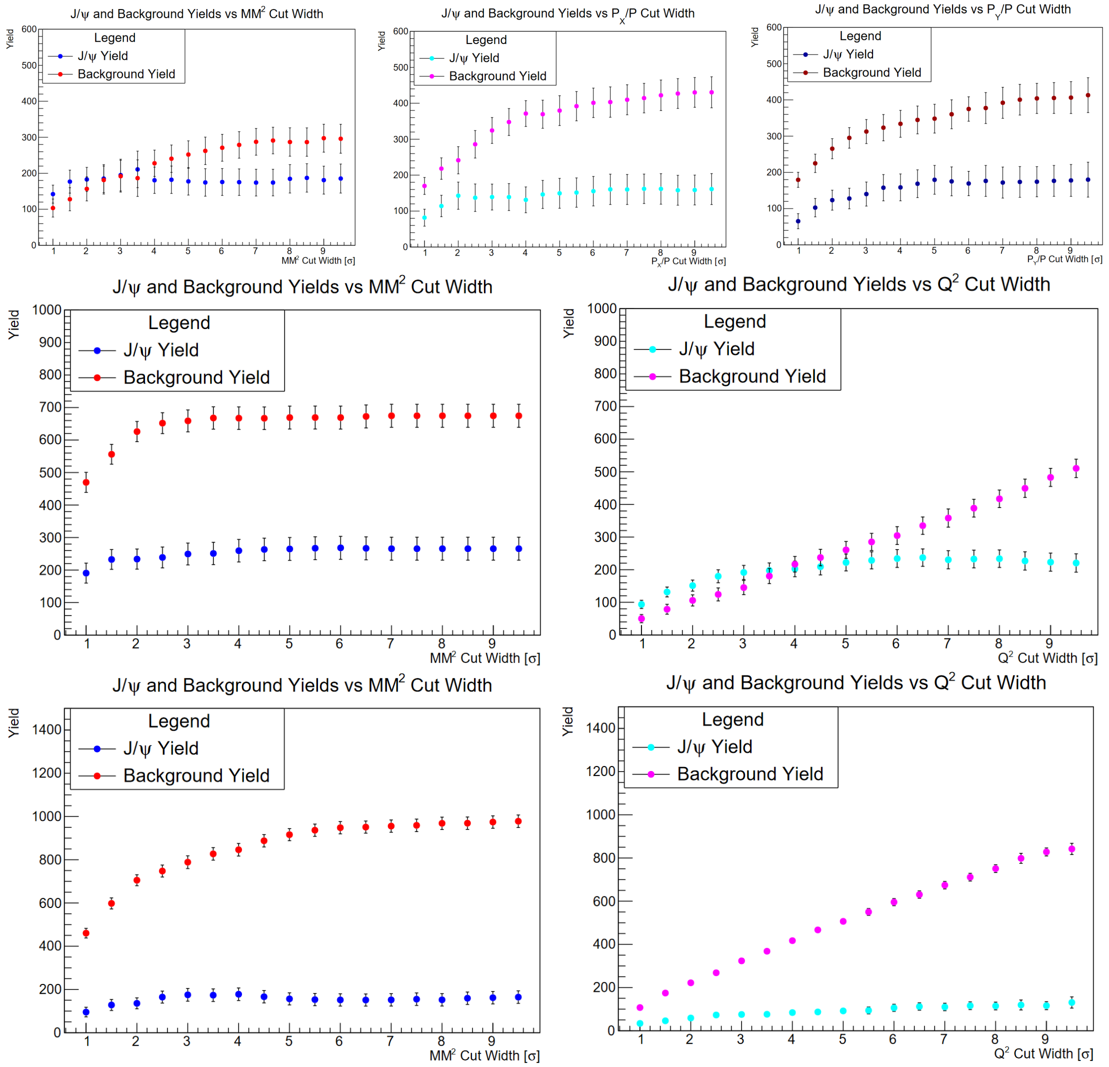


Figure 7.7: The J/ψ and background yields as a function of the cut width on the missing mass squared (Left) and Q^2 (Right), for the channels $ep_{bound} \rightarrow (e')e^+e^-p$ (middle row) and $en_{bound} \rightarrow (e')e^+e^-n$ (bottom row), and on the missing mass squared (Left), $\frac{P_X}{P}$ (Center) and $\frac{P_Y}{P}$ (Right) for the channel $ep \rightarrow (e')\mu^+\mu^-p$ (top row). The cut width is here counted in units of the σ returned by the parametrisations as a function of E_γ.

duction events are a subset. In this case the training sample is composed of a J/ψ simulation for the positive sample, with the negative sample a combitorial background created by mismatching particles from different events of the J/ψ simulation. As shown in Figure 7.9, this ensures that the kinematics of individual particles are roughly the same in both the positive and negative sample. However, as shown in Figure 7.8, the missing mass squared and the di-lepton invariant mass distributions are different. The learning algorithm will therefore be trained to correlate the kinematics of the individual particles so as to learn to recognise quantities such as the missing mass squared, and the correlation of these quantities and their resolution to the individual particle kinematics.

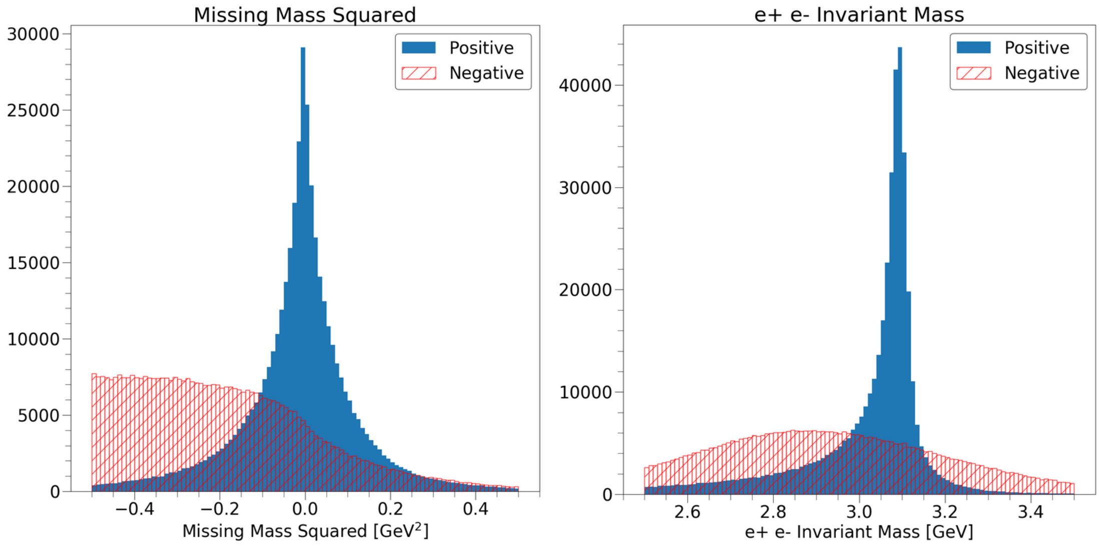


Figure 7.8: The missing mass squared (Left) and di-lepton invariant mass (Right) for the positive and negative samples of the training data set.

This machine learning approach to event selection was tested for J/ψ photoproduction on the bound proton, with J/ψ further decaying to an electron positron pair. The training samples are composed of the four momentum of the final state particles, as shown in Figure 7.9. The four-momentum distributions are roughly the same for the negative and positive samples, ensuring that the learning algorithm will not learn the individual particle kinematics but that of the event as a whole.

For this use case, a neural network was trained in *tensorflow*. The neural network was composed of five layers, the input layers having nine nodes and a ReLu activation function, the last layer have two nodes and a softmax activation function, with more detail on these activation functions found in Chapter 5. The three inner layers had 56, 28 and 14 nodes, each with a ReLu activation function. The neural network used an ADAM optimiser with a learning rate of 10^{-4} and a binary cross entropy loss function. The neural network was trained for 50 epochs. Figure 7.10 shows the loss evaluated on the training at testing set at each epoch. There is no notable difference between the loss on the training or testing sets indicating that the neural network is not overfitting.

Once trained, the neural network was applied to the testing set to evaluate metrics such as the purity, accuracy, efficiency. The response, and the metrics for various thresholds on the response are shown in Figure 7.11. Also shown is the the invariant mass distribution of events rejected and selected by at cut at 0.5 on the response. These are further separated into true negatives and false negatives, which are negative events correctly rejected and positive events incorrectly rejected, and true and false positives which are positive events correctly selected and negative events incorrectly selected. As shown, positive events that are incorrectly rejected occur

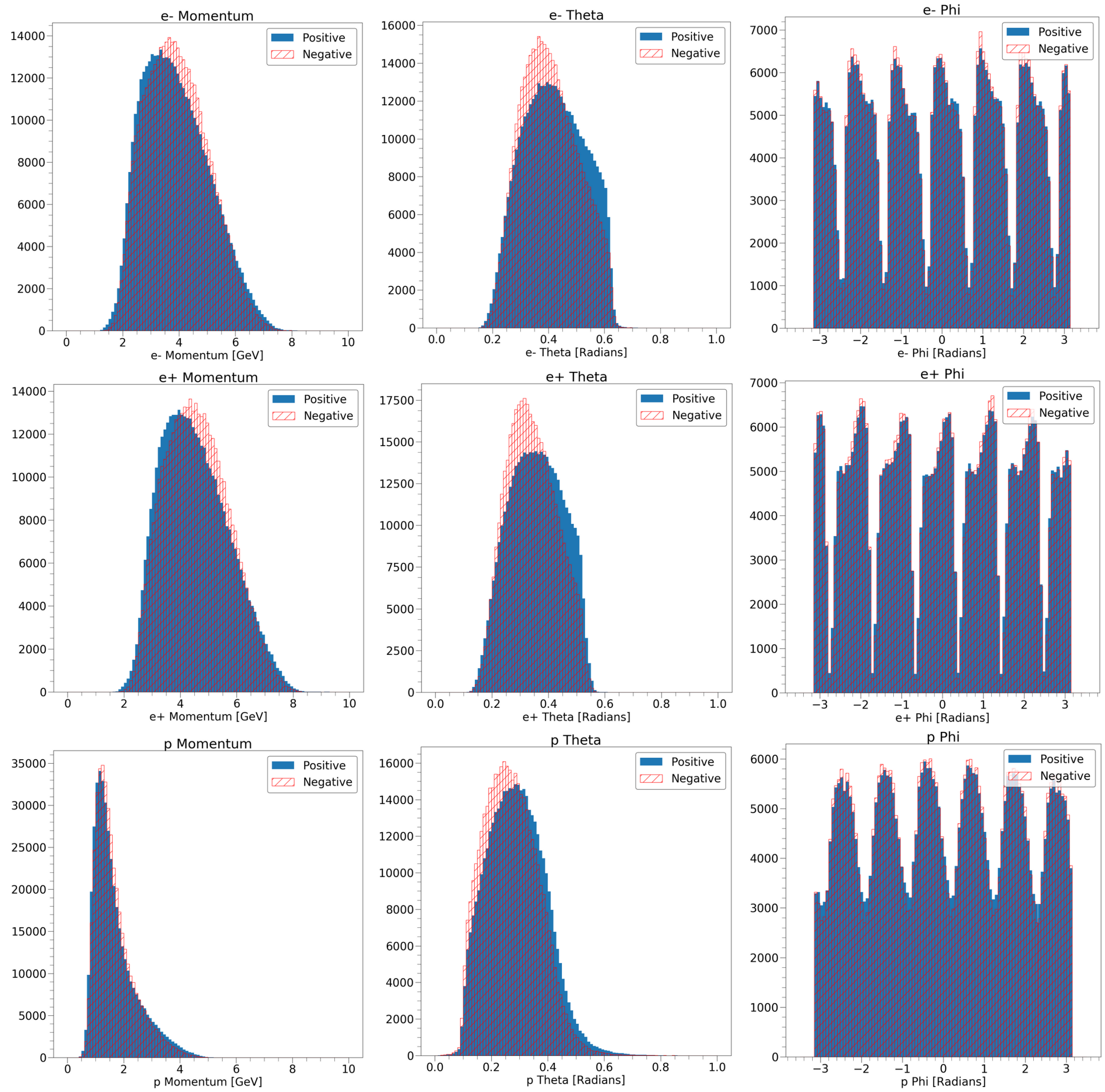


Figure 7.9: The momentum (Left), polar (center) and azimuthal (Right) angles of the electron (top row), positron (center row) and proton (bottom row) for the positive and negative samples of the training data set.

predominantly in the radiative tails of the J/ψ peak, although some events at the peak were also lost by the neural network. However, the neural network is capable of selecting events in the radiative tails, which suggests that it does not simply select events that have an invariant mass close to the J/ψ mass. Figure 7.11 also shows the invariant mass distribution of the predicted positives and negatives compared to the true positives and negatives (center). The predicted positives are composed of both true and false positives, with the predicted negatives composed of both true and false negatives. A dip in the predicted negatives compared to the true negatives is observed close to the J/ψ peak, which is accompanied by a slight increase in the height and width of the J/ψ peak. Already these are indications that the neural network is somewhat biased towards the J/ψ mass, even on the simulated testing data.

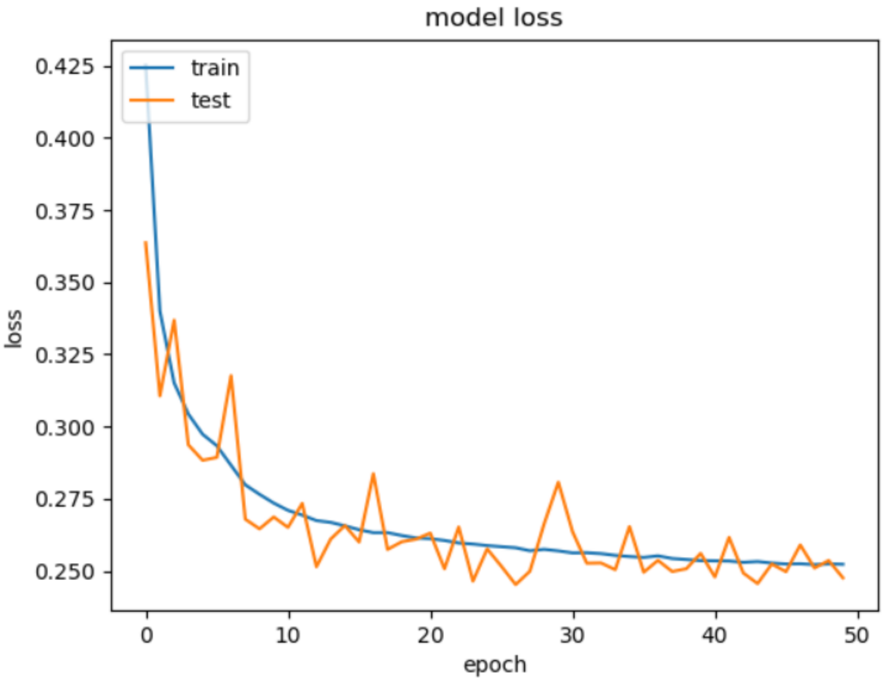


Figure 7.10: The binary cross entropy loss evaluated at each epoch on the training and testing sets.

The next step is to try out the neural network on data taken with CLAS12. Figure 7.12 shows the J/ψ and background yields as a function of the threshold on the response for J/ψ photoproduction data at CLAS12. Events were selected with a cut on Q^2 at 0.5 and a positron identification cut of 0.3 on the positron ID response. A gaussian function was used to fit to the J/ψ peak, with a parameter of the fit returning the J/ψ yield, and a third order polynomial was fitted to the background. Although the background is significantly decreased, the J/ψ yield is increased for a response cut between 0.5 and 0.95. This is worrying as it is nonphysical; the neural network should not find additional J/ψ, only remove these and background. Similarly, the σ and mean of the gaussian fit is plotted at each threshold on the response in Figure 7.12. This shows how the σ increases as the threshold is increased, before decreasing once the threshold increases past 0.9. This is problematic as it suggests that the neural network is altering the fit hypothesis, which essentially betrays the fact that the neural network is biased towards the J/ψ mass.

Figure 7.13 shows the di-lepton invariant mass spectrum at a threshold on the response of 0 (which is akin to not using the neural network), of 0.86 and of 0.94. Clearly at a response of 0.86 the width of the gaussian fitted to the J/ψ peak is increased which leads to an increase in the J/ψ yield. This is problematic when, for example, measuring cross sections as this would artificially increase the cross section. However, at a response of 0.94, the fit parameters are consistent with those returned without using the neural network. The background is also notably suppressed. A suggestion would therefore be to only use the neural network in conjunction with a high cut on

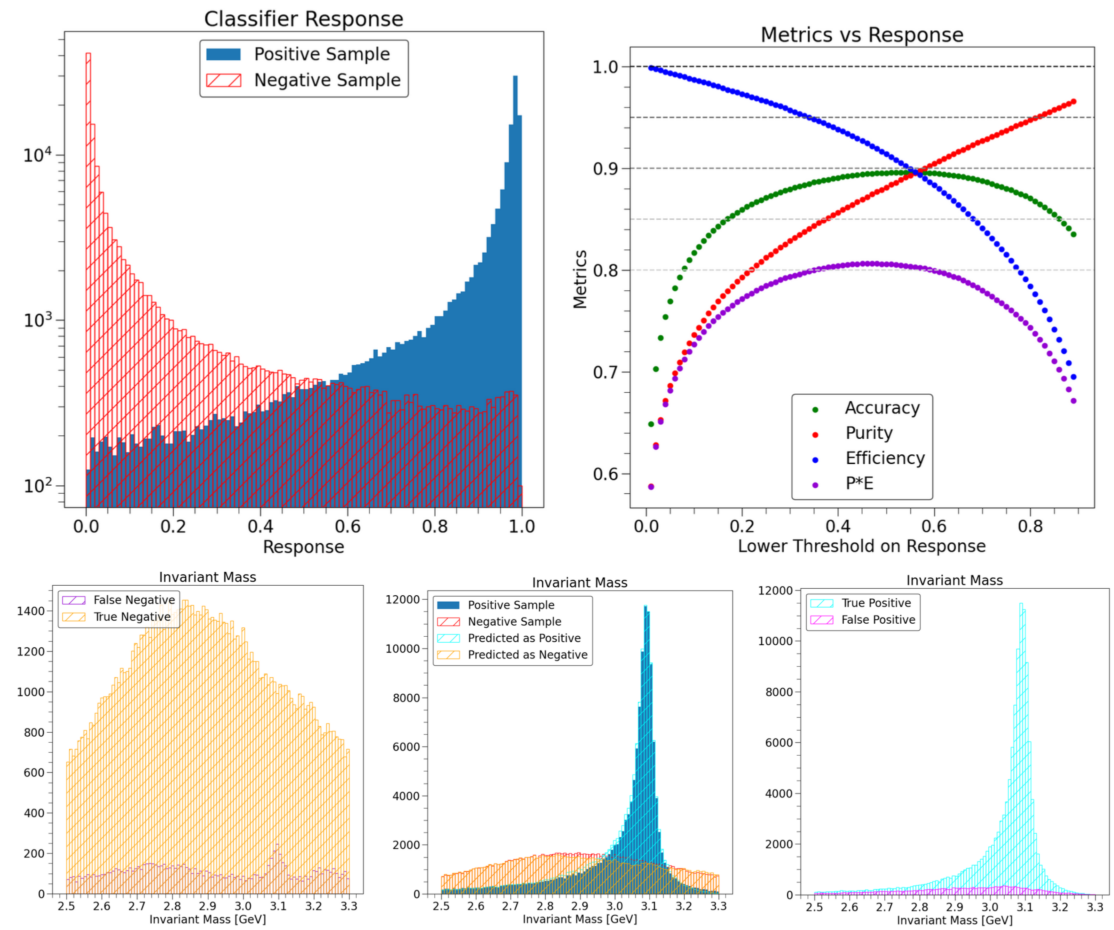


Figure 7.11: Top Row: The Response (Left) of the neural network classifier on the positive (blue) and negative (red) samples of the testing set, as well as metrics evaluated at various thresholds on the response (Right). Bottom Row: The invariant mass distribution of events rejected (left) and selected (right) by at cut at 0.5 on the response, and the invariant mass distribution of the predicted positives and negatives compared to the true positives and negatives (center).

the response.

Overall, great care should be taken when using machine learning based methods like the one described in this section. Figures 7.11, 7.12 and 7.13 demonstrate clear evidence of a bias introduced in the analysis by this method, which would directly impact the extraction of physics results such as the cross section. Chapter 8 will show a comparison of the cross section when using or not the neural network and at different cuts on the response. The suggestion here not being that the cross section should be measured with such a neural network but rather out of curiosity to judge how the cross section is affected.

However, as shown in Figures 7.12 and 7.13, the background is greatly suppressed by the neural network. The network also has a high prediction rate of order $10^5 s^{-1}$. Reaction identification tools such as the one developed here can still be useful for different applications. Machine

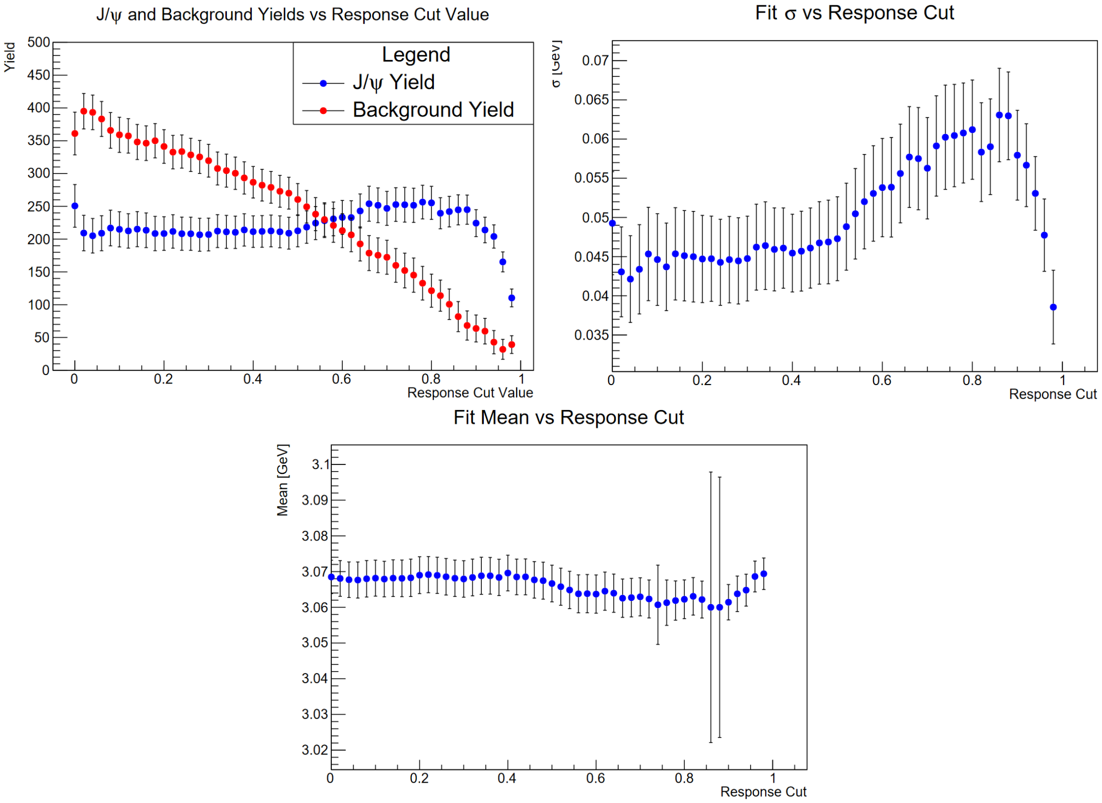


Figure 7.12: Top Left: The J/ψ and background yields as a function of the threshold on the response evaluated on CLAS12 data. Top Right: The σ of the gaussian fitted to the invariant mass distribution at each threshold on the response. Bottom: The mean of the gaussian fitted to the invariant mass distribution at each threshold on the response.

learning based tools have been shown to be able to predict the four momentum components of charged tracks from hits in the CLAS12 drift chambers with a reasonable resolution [145]. By combining this approach to momentum reconstruction to the tools described in this section, fast reaction identification could be achieved at the online stage during data taking. This could then allow to do online calibrations dependent on identifying certain reactions. Yields when measured with a high cut on the response, and taken with a pinch of salt, could also be used for online monitoring purposes. Online flagging of certain reactions could then allow for shorter timescales in producing physics measurements as only specific events could be targeted for reconstruction, removing an important timing overhead due to reconstructing background events, with more detailed analyses allowing to correct the bias introduced by the reaction identification tools. Appendix A.4 describes the first initial steps that could be taken towards a near-online analysis at CLAS12. The reaction identification tools may also be useful for measurements relying on asymmetries where the difference of two observables is divided by their sum, as any increase in the observables would factor out. All the suggestions above however remain to be investigated but they open up an exciting prospect for online analysis.

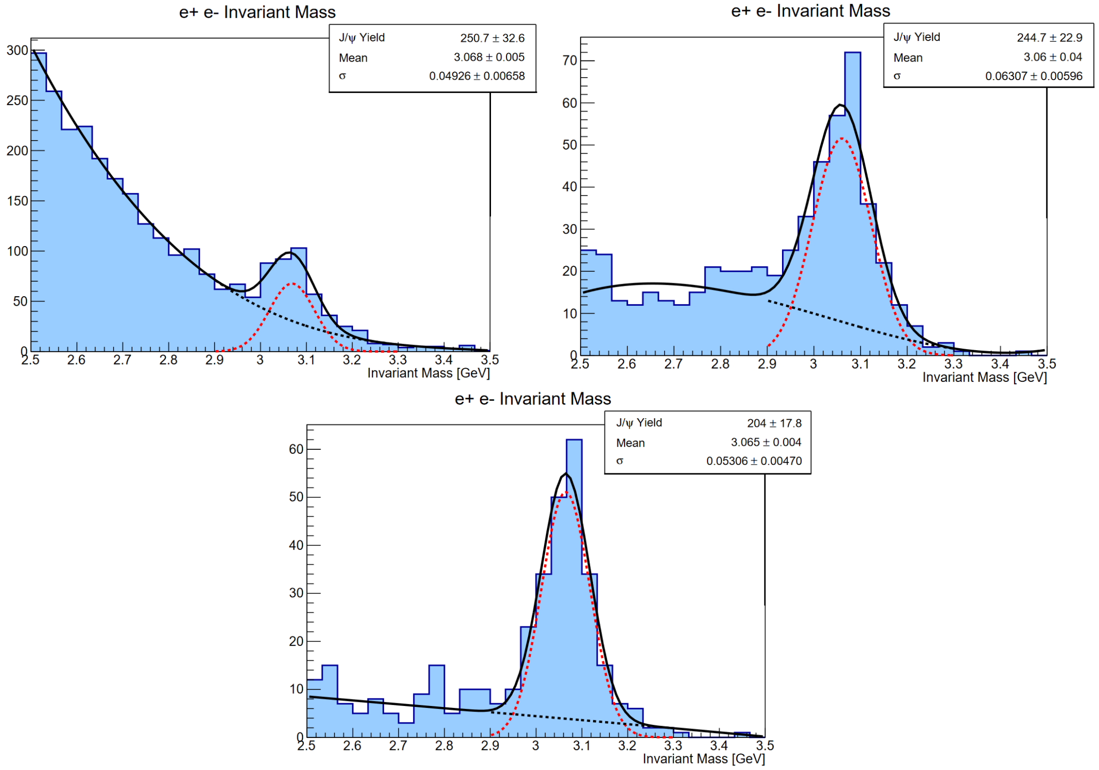


Figure 7.13: The di-lepton invariant mass distribution at a threshold on the response of 0, 0.86 and 0.94. A threshold of 0 is akin to not using the neural network. A gaussian is fitted to the J/ψ peak with a third order polynomial fitted to the background.

7.3 The Di-lepton Invariant Mass

Once J/ψ photoproduction events have been identified, the last step of the analysis before calculating the J/ψ photoproduction cross section is to measure the number of J/ψ produced in a given data set. The di-lepton invariant mass calculated using Equation 2.4 is then fitted to extract the number of J/ψ. Although this has been shown in this chapter and the previous one when considering how certain analysis procedures affect the J/ψ yield, it is worth spending a little bit of time considering some of the main features of the invariant mass spectrum.

Figure 7.14 shows the di-lepton invariant mass for an electron positron pair produced on a bound proton or a bound neutron, and for a di-muon pair produced on a proton. For the channel $ep_{bound} \rightarrow (e')e^+e^-p$, the electron and proton were selected with the EBPID, the positron was selected with the EBPID and the identification was refined with a cut on the positron ID response at 0.1 with a cut at 2.5 σ and 3 σ on the missing mass squared and Q^2 parametrisations respectively. The Q^2 parametrisation was established on simulated data. For the channel $en_{bound} \rightarrow (e')e^+e^-n$, the electron was selected with the EBPID, the positron was selected with the EBPID and the identification was refined with a cut on the positron ID response at 0.3 with a cut at 3 σ and 10 σ on the missing mass squared and Q^2 parametrisations respectively. The cut at 10 σ on the Q^2 parametrisation is deliberately loose to remove some electroproduction background without having any effect on the J/ψ yield. The neutron was detected, identified and corrected as explained in Chapter 6. For the $ep \rightarrow (e')\mu^+\mu^- p$ channel, the di-muon pair was selected using the energy deposition cuts described in Chapter 6 and with a cut on the muon ID response at -0.12. A cut was also placed at 2.5 σ on each of the missing mass squared, $\frac{P_X}{P}$ and $\frac{P_Y}{P}$ parametrisations.

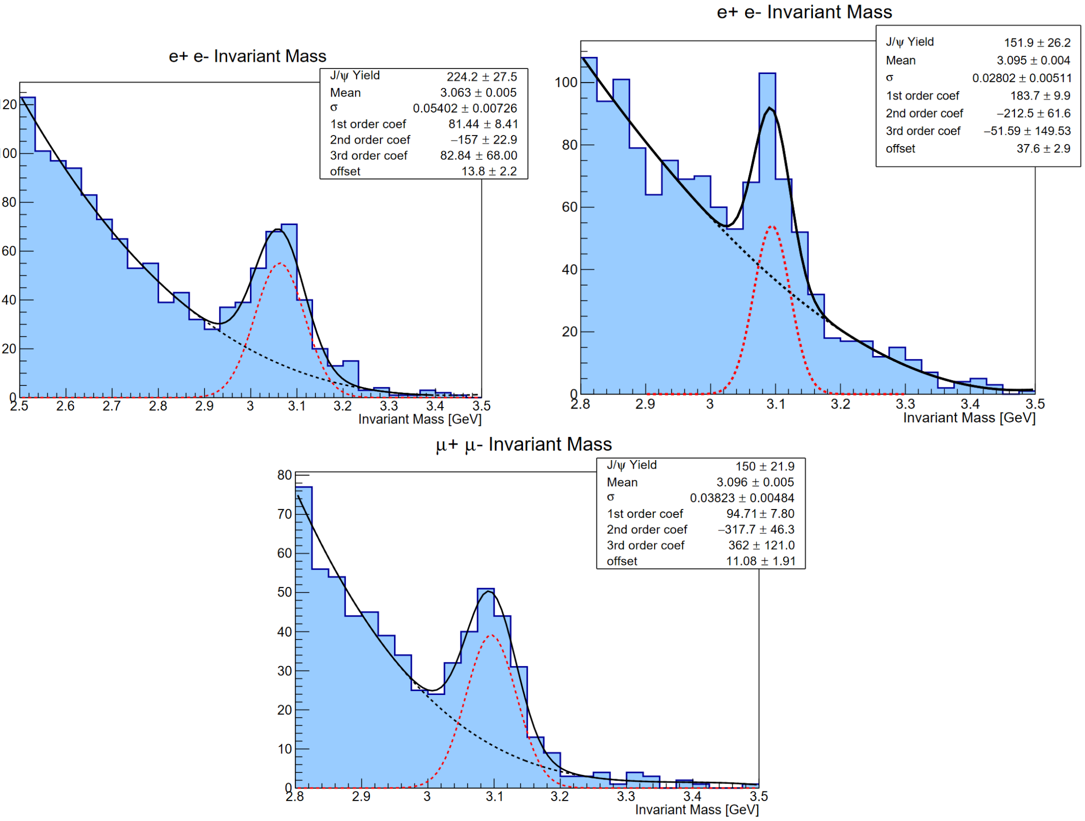


Figure 7.14: The di-lepton invariant mass distribution in the channels $ep_{bound} \rightarrow (e')e^+e^- p$ (Top Left), $en_{bound} \rightarrow (e')e^+e^- n$ (Top Right) and $ep \rightarrow (e')\mu^+\mu^- p$ (Bottom).

As before, the invariant mass spectra of Figure 7.14 were fitted with a gaussian for the J/ψ peak and a third order polynomial for the background. A parameter of the gaussian fit returns the J/ψ yield. Each plot of Figure 7.14 gives the parameters of the fit functions. The J/ψ yields are slightly different, due to the datasets on a deuteron and hydrogen target taken over different periods of time, and due to varying acceptance in the different channels. However, interestingly, the width and mean of the fits are significantly different.

Figure 7.15 shows the di-muon invariant mass spectrum for J/ψ produced on a bound proton or bound neutron, decaying to a di-muon pair. The $ep_{bound} \rightarrow (e')\mu^+\mu^- p$ requires a di-muon pair with the muon ID response cut placed at 0.08, and straight cuts on the missing mass squared and Q^2 at 0.3 GeV^2. The channel $en_{bound} \rightarrow (e')\mu^+\mu^- n$ requires a di-muon pair with the muon

ID response cut placed at 0.15, and uses the same missing mass squared parametrisation as that established for the $en_{bound} \rightarrow (e')e^+e^-n$, with a cut placed at 5σ. A cut on Q^2 was placed at 1 GeV. These channels are considered here for comparison to the channels shown in Figure 7.14, but will not be analysed further in this thesis due to low statistics and time constraints. As explained in Chapter 4, a pre-skim was made of CLAS12 data to pre-select events suitable for J/ψ photoproduction. This pre-skim contained energy deposition requirements for the di-muon pair that were too restrictive and decreased the J/ψ yield. Another pass over this data will fix this issue, along with generally providing more statistics. The fit to the invariant mass spectra offers interesting comparisons in terms of the mean and σ of the gaussian fitted to the J/ψ peak.

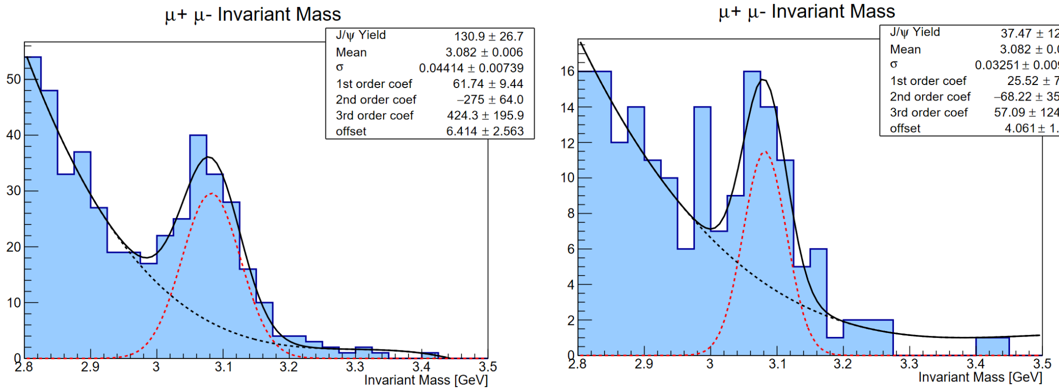


Figure 7.15: The di-lepton invariant mass distribution in the channels $ep_{bound} \rightarrow (e')\mu^+\mu^-p$ (Left) and $en_{bound} \rightarrow (e')\mu^+\mu^-n$ (Right).

Figure 7.16 shows a comparison of the mean and sigma of the gaussian fitted to the J/ψ peak in the invariant mass spectra showed in Figures 7.15 and 7.14. The first thing to note is a roughly 15 MeV shift in the mean when comparing the di-muon channel produced on a proton with those produced on a bound proton or neutron target, and when comparing the di-electron channel produced on a proton with that produced on a bound proton. The results for the di-electron channel produced on a proton were given by CLAS12 collaborators [84]. This suggests an overall issue with the calibration of the dataset taken on a deuteron target which shifts the mean by roughly 15 MeV. However, more concerning is the fact that the J/ψ peak mean is different by about 30 MeV in the di-electron channel produced on a bound proton with the di-electron channel produced on a bound neutron. The invariant mass is calculated from the di-lepton pair, and so whether J/ψ is produced on a proton or neutron should not affect the peak mean. Similarly, the σ of the fit is consistent between all di-muon channels, and between the di-electron channels produced on a proton and bound proton. However, the σ for J/ψ produced on a neutron is smaller by roughly 15 MeV than when produced on a bound proton or proton. Again, there is no a priori reason why that should be the case. One would expect the consistency observed in the di-muon channels. Note that a smaller σ on average in the di-muon channels compared to the di-electron channels is expected as muons will not radiate photons, as mentioned in Chapter 6.

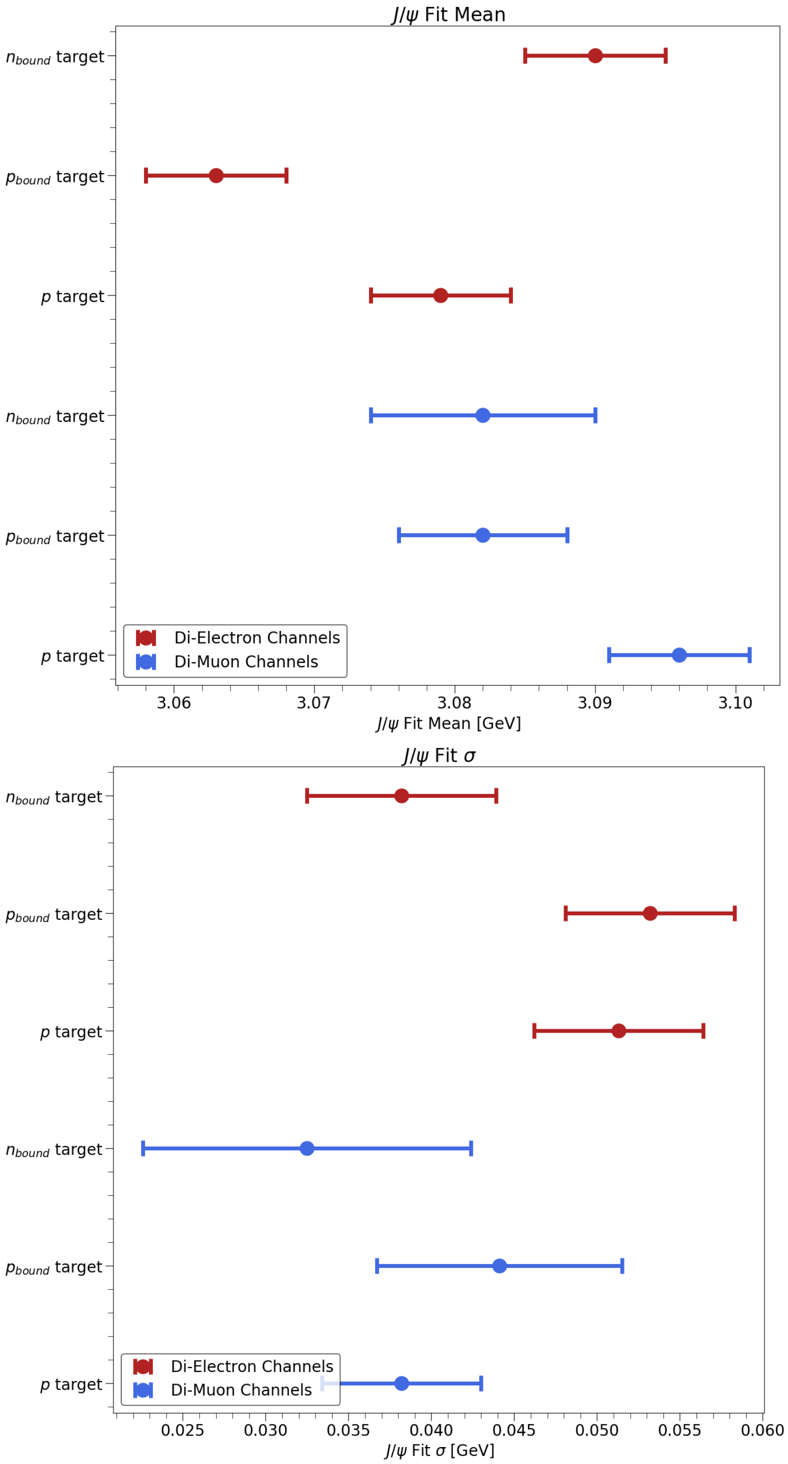


Figure 7.16: The mean and σ of the gaussian fitted to the J/ψ peak on proton, bound proton or bound neutron target, for J/ψ decaying to an electron positron (red) or di-muon (blue) pair.

The invariant mass spectra above were measured with very low statistics, especially so for the channels produced on the neutron. Two types of tests were made to assess the statistical significance of the observed differences. The first generated 10^4 gaussians with 100 events and a mean of 3.097 GeV and σ of 0.05 GeV. These were then fitted with the same gaussian as the CLAS12 J/ψ photoproduction data, and the mean and sigma of each is plotted in histograms in Figure 7.17. As shown, the σ on the mean is roughly 6 MeV, and that on the σ of the J/ψ fit is roughly 5 MeV. This seems to suggest that a difference on the mean of roughly 30 MeV and a difference in σ of roughly 15 MeV are not due simply to statistical effects.

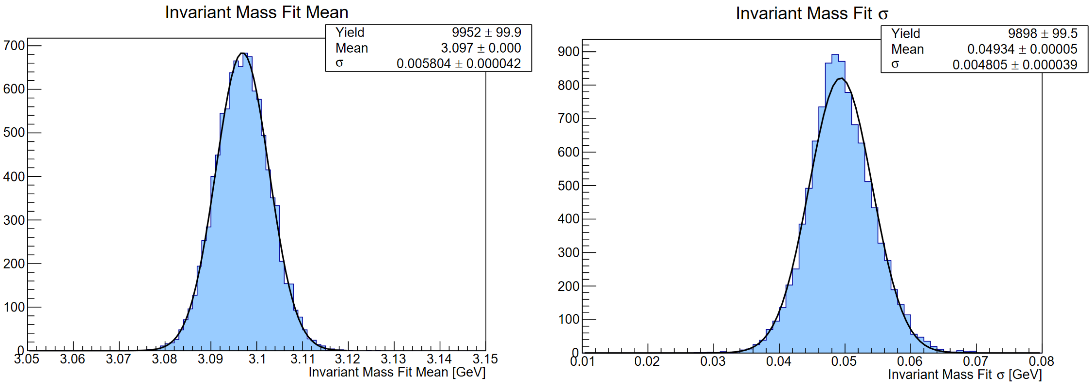


Figure 7.17: The mean and σ of the gaussian fitted to 10^4 gaussian distributions with 100 events and a mean of 3.097 GeV and σ of 0.05 GeV.

The second set of tests used the T-test to assess if the means of two normal distributions are significantly different and Levene's test to assess the equality of variances of two normal distributions. Both tests return a p-value. If the p-value is under a threshold, typically 0.05, the differences are unlikely to have occurred due to random sampling from the same distribution. For both tests a gaussian distribution with 100 events is generated with the mean and σ measured for J/ψ produced on a bound proton, decaying to an electron positron pair. A second gaussian distribution with 100 events is generated with either the same fixed mean or σ but varying the unfixed parameter. The T-test is then produced by comparing the first gaussian distribution to the one with varying mean, and the Levene's test is compared between the first gaussian distribution to the one with varying σ. The results of the tests, averaged over 1000 repetitions at each point, are plotted in Figure 7.18. Once again, the results of the tests seem to suggest that it is unlikely that the gaussian fits to the bound proton and bound neutron channels shown in Figure 7.14 occurred due to random sampling from the same distribution.

An important caveat to the tests shown in Figures 7.17 and 7.18 is that they assume that the J/ψ peak is perfectly gaussian. In practice it is more likely to resemble a crystal ball function where the gaussian peak presents long asymmetrical tails, which are due to the electron and positron radiating photons and therefore loosing some momentum. Furthermore, they also assumed that the background fit hypothesis does not affect the gaussian fit to the J/ψ peak. In practice this is not the case, and might lead to variations in the recorded mean and σ. A

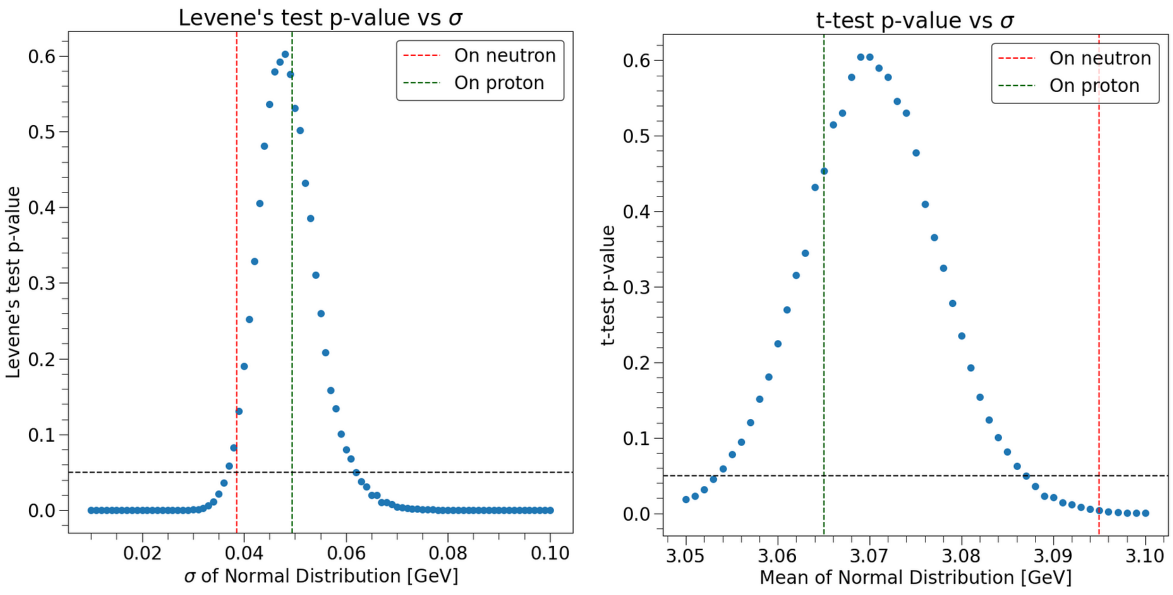


Figure 7.18: The results of Levene's test (Left) and the T-test (Right) averaged over 10^3 iterations as a function of the varying mean or σ.

more detailed study is required to establish more realistic fit functions to the J/ψ peak and the background. Given that a new pass of CLAS12 data will produce more statistics, as detailed in Chapter 4, it seemed like a better use of time to wait for better statistical precision before doing this fitting study.

Another consideration when looking at the di-lepton invariant mass spectra is that, although Figures 7.14 and 7.15 were zoomed in around the J/ψ mass, other resonances are clearly identifiable when looking at the invariant mass on its full range. Figure 7.19 shows the full invariant mass range for the $ep_{bound} \rightarrow (e')e^+e^-p$ channel, with the same cuts as in Figure 7.14. The peak at a mass of zero is due to the Dalitz decay of photons to an electron positron pair. The peak at 0.12 is due to electrons and positrons from different events that can be cut out by requiring a time difference between the two less than the width of the beam time bunching. The peak at roughly 0.75 GeV is due to the ρ and ω resonances, that are here indistinguishable. Finally, the ϕ resonance is clearly visible and here fitted with a gaussian, with the background in the region of the ϕ mass fitted with a third order polynomial. As mentioned in Chapter 2, there is some suggestion in literature that ϕ photoproduction could be used to measure similar physics quantities as those observed in J/ψ photoproduction. As such, the analyses of the $ep \rightarrow (e')e^+e^-p$ and $ep_{bound} \rightarrow (e')e^+e^-p$ channels can be used both to measure J/ψ and ϕ photoproduction at CLAS12, although the main focus of this thesis remains J/ψ photoproduction. Note that the same cannot be said for the di-muon analyses, due to a very large background of ρ decaying to two pions that washes out the ϕ decaying to two muon peak. The ϕ analyses also require the recoil nucleon to be detected in the central detector, which at the moment presents some issues with neutron reconstruction. As such ϕ photoproduction at CLAS12 has for now only been measured on the proton or bound proton, with ϕ decaying to an electron positron pair.

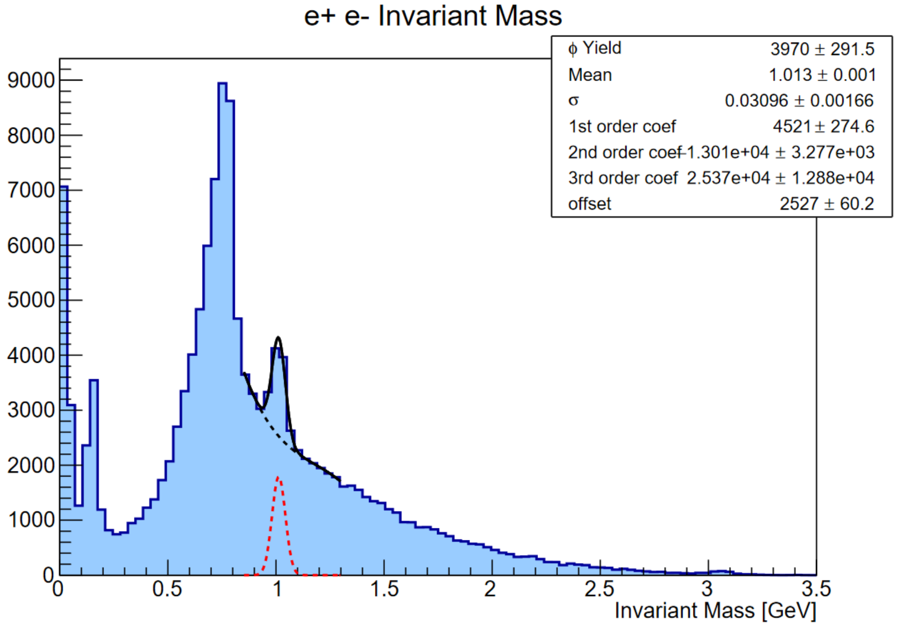


Figure 7.19: The di-lepton invariant mass distribution in the channel $ep_{bound} \rightarrow (e')e^+e^-p$. The ϕ resonance is here fitted with a gaussian, with the background in the region of the ϕ mass fitted with a third order polynomial.

The invariant mass of all decay products, W, which is equal to the invariant mass of $J/\psi p$, is shown in Figure 7.20 for the channel $ep_{bound} \rightarrow (e')e^+e^-p$. This is firstly shown when using cuts on the missing mass squared and Q^2 in the di-lepton invariant mass range of 2.9 to 3.1 GeV. Structure around 4.3 GeV or 4.5 GeV would hint to the existence of the P_C^+ pentaquarks. No clear evidence for this is observed in Figure 7.20. W is also plotted using cuts on Q^2 and the reaction identification tool to remove as much background as possible, with the same requirements on the di-lepton invariant mass range. This did not yield any additional evidence for the existence of the P_c^+ pentaquarks.

7.4 Non-Exclusive Backgrounds

The cross sections that this thesis aims to measure are that of exclusive J/ψ photoproduction, meaning that only J/ψ is produced in the interaction of the electron beam with the target. To avoid making any assumptions on the relative cross sections of exclusive to non-exclusive J/ψ photoproduction, two non-exclusive channels were simulated for J/ψ produced on a hydrogen target and decaying to a di-muon pair:

$$ep \rightarrow e'J/\psi\pi^0 p$$

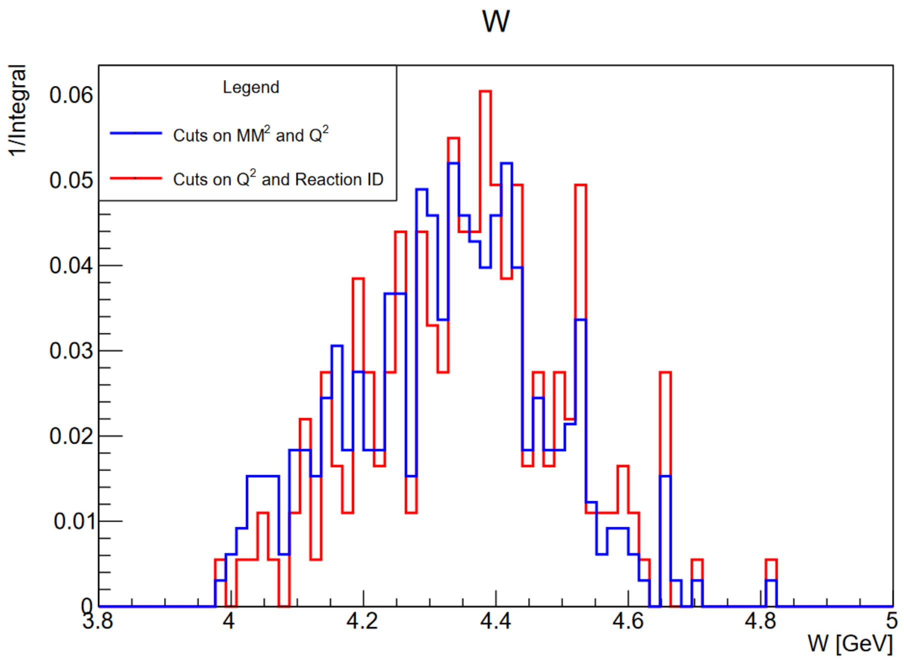


Figure 7.20: The $J/\psi p$ invariant mass, W, in the channel $ep_{bound} \to (e')e^+e^- p$ compared using only cuts on the missing mass squared (MM^2) and Q^2 in blue to using cuts on Q^2 and the reaction identification tool in red. The invariant mass of the lepton pair is restricted between 2.9 and 3.1 GeV. Both histograms are scaled by their integral for ease of comparison. The fine binning is used to emphasise any possible structure.

$$ep \to e'J/\psi\Delta \to e'J/\psi\pi^0 p$$

Additional non-exclusive channels may exist, but these require the least additional energy to produce and should therefore be the most kinematically favourable. Any evidence for these channels would warrant more in depth analysis of non-exclusive backgrounds. These were also simulated for ϕ photoproduction, where ϕ decays to an electron-positron pair, as the increase in statistics may help identify non-exclusive backgrounds. Shown in Figure 7.21 is the missing mass squared of the simulated exclusive and non-exclusive channels for both J/ψ and ϕ photoproduction. No cuts on the missing mass squared or Q^2 are applied here. The non-exclusive channels present a shoulder at positive missing masses due to the fact that the neutral charged pion goes undetected in the usual analysis that only requires the recoil nucleon and the decay leptons of J/ψ. The shoulder for the non-exclusive J/ψ channels is more peak like than for ϕ due to the fact that J/ψ requires more energy than ϕ and so less energy can be carried by the missing pion.

Figure 7.21 assumes that the exclusive and non-exclusive channels have equal cross sections but this most likely is not the case. To investigate this, the invariant mass of the di-muon pair or electron-positron pair is fitted in the vicinity of the J/ψ or ϕ mass respectively using the *brufit* [147] package. Using the sPlot method [148, 149], weights are assigned to every event,

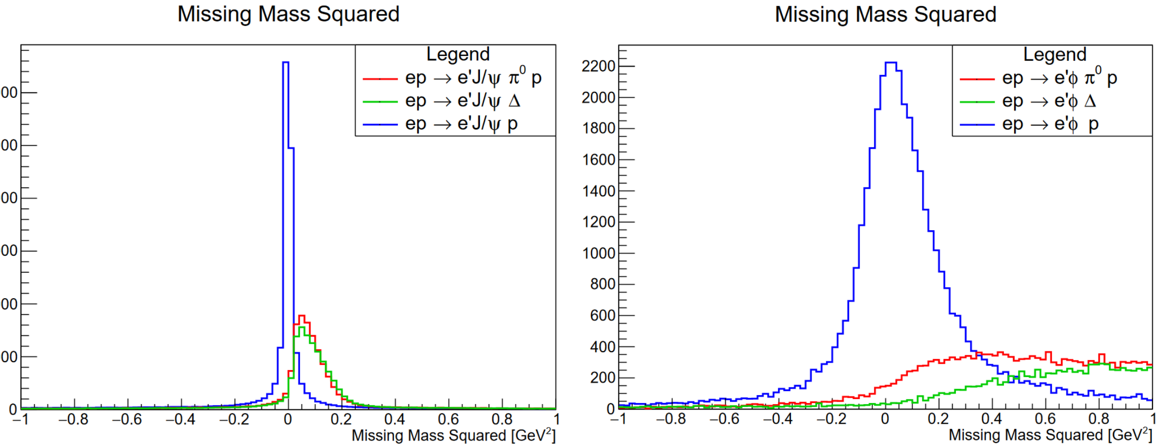


Figure 7.21: The missing mass squared of simulated exclusive and non-exclusive channels for J/ψ (left) and ϕ (right) photoproduction.

allowing to distinguish between the signal and background distributions fitted to the data. The signal, in this case J/ψ or ϕ, was fitted with a gaussian. The background was fitted with a second order Chebychev polynomial for J/ψ and with a third order Chebychev polynomial for ϕ. Cuts on Q^2 and the missing mass squared at 0.7 GeV^2 were applied to both channels to help fit the signal peak, with an additional tight cut on the di-muon machine learning classifier helping to fit the J/ψ peak. These cuts are sensible to estimate the non-exclusive background contributions as they would typically be applied for the measurements of the cross sections. The resulting fits and fit parameters are shown in Figure 7.22.

The weights assigned to each event then allow to only preserve the signal events, in this case J/ψ or ϕ events. The missing mass squared for these two sets of events can then be fitted using the simulated exclusive and non-exclusive channels. Three parameters, alpha, off, scale, allow to convolute the simulated models to represent the data. The parameter off allows to offset the mean of the distribution, with the scale parameter scaling the width. The parameter alpha allows to convolve the simulated model with an additional gaussian to better fit the data. Figure 7.23 shows the resulting fits for the J/ψ or ϕ events. No evidence for the non-exclusive backgrounds is found in the missing mass squared distributions as the fit assigned a yield of zero to the non-exclusive channels. This essentially confirms that the analyses presented in this thesis are for exclusive J/ψ and ϕ photoproduction.

7.5 Measuring the J/ψ SDMEs

As mentioned in Chapter 2, SDMEs fully describe the angular distribution of a meson's decay. Measuring the polarised SDMEs would allow to distinguish between various production mechanisms, such as two- or three-gluon exchange. In addition, including the SDMEs in the

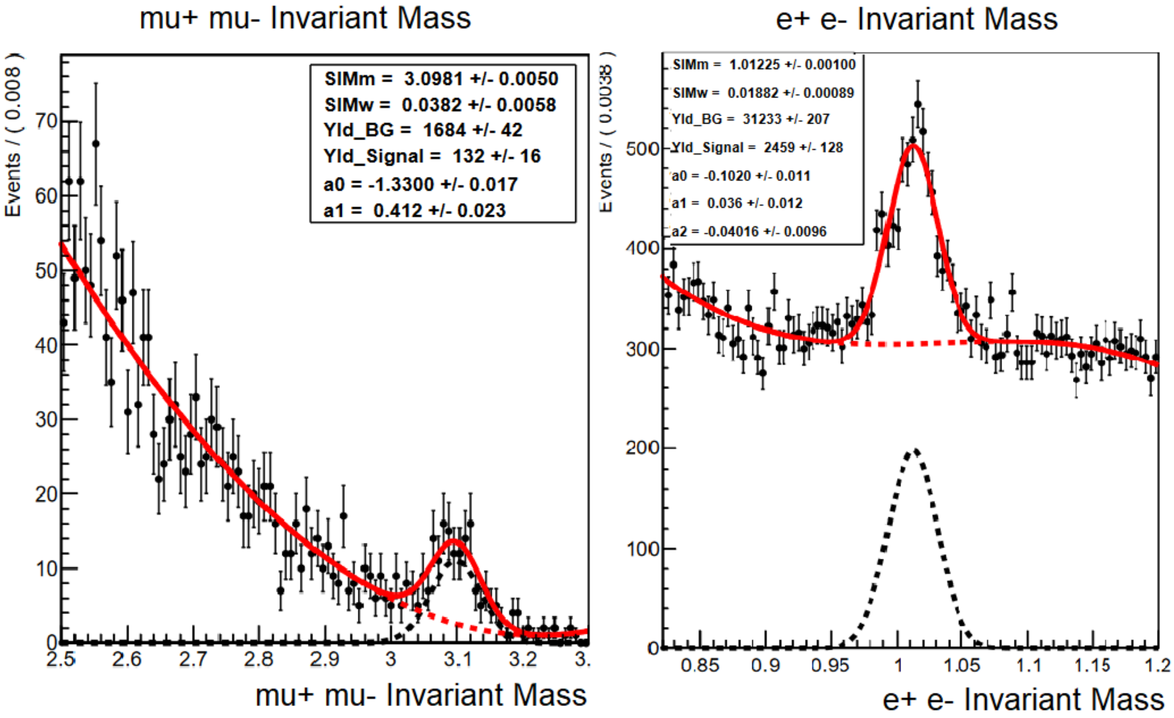


Figure 7.22: The di-lepton invariant mass fitted with a gaussian and polynomial background in the region of the J/ψ mass (left) or ϕ mass (right) using the sPlot technique.

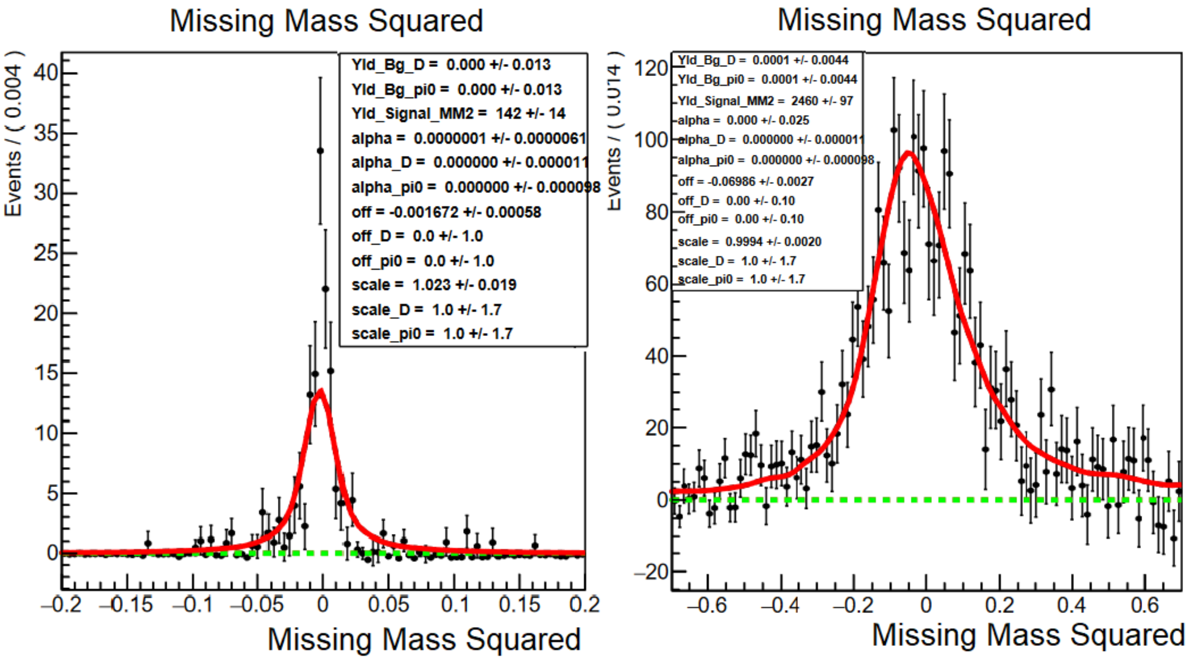


Figure 7.23: The missing mass squared of events identified as J/ψ (left) or ϕ (right) events using the sPlot weighting technique. No evidence of non-exclusive backgrounds is found in the missing mass squared distributions as indicated by the essentially zero yields returned by Yld_Bg_pi0 and Yld_Bg_D.

simulations would allow for a better representation of J/ψ photoproduction at CLAS12 as the angular distribution of the decay products of J/ψ would be better simulated. As the detection efficiency of the CLAS12 detector might vary as function of polar or azimuthal angle, encoding the SDMEs into the simulation means that the acceptance calculation would be more realistic.

For now only the unpolarised SDMEs ρ^0_{00}, ρ^0_{10}, and ρ^0_{1-1} were measured with CLAS12 given the low statistics available. These are measured by fitting the decay angles in the Gottfried Jackson frame, θ_{GJ} and ϕ_{GJ}. The Gottfried Jackson frame is measured in the rest frame of J/ψ with the z-axis taken as the direction of the incident quasi-real photon and naturally describes the angular distribution of the J/ψ decay. The decay angles θ_{GJ} and ϕ_{GJ} are fitted with the following equation [78]:

$$\begin{aligned} W(\cos(\theta_{GJ}),\phi_{GJ}) =& \frac{3}{8\pi}\Big(\frac{1+\rho^0_{00}}{2} - \frac{3\rho^0_{00}-1}{2}\cos^2(\theta_{GJ}) + \sqrt{2}Re\rho^0_{10}\sin(2\theta_{GJ})\cos(\phi_{GJ}) \\ &+ Re\rho^0_{1-1}\sin^2(\theta_{GJ})\cos(2\phi_{GJ})\Big) \end{aligned} \tag{7.2}$$

To this end, the *brufit* [147] package was used first to fit the electron positron invariant mass in the $ep_{bound} \to (e')e^+e^-p$ channel. Using the sPlot method [148, 149], weights are assigned to every event, allowing to distinguish between the signal and background distributions fitted to the data. The background was fitted with a third order Chebychev polynomial. The signal was fitted using a simulated model. Three parameters, alpha, off, scale, allow to convolute the simulated model to represent the fitted data as detailed in the previous section. Figure 7.24 shows the fit to the invariant mass.

The signal events are then identified by the sPlot weights assigned to each event. The decay angles θ_{GJ} and ϕ_{GJ} are then fitted using Equation 7.2 with *brufit* [147]. To be precise, the cosine of θ_{GJ} is fitted here. The resulting fits are shown in Figure 7.25.

The SDMEs can also be measured in the same way using simulated data that was generated using the Joint Physics Analysis Center (JPAC) model for the J/ψ near-threshold cross section, as mentioned in Chapter 4. These are not shown here due to a very large statistical uncertainty on the measured SDMEs, and some disagreement with the JPAC model. However, the measurements along with the JPAC predictions can then be used to weight simulated events using Equation 7.2 to check for any systematic effect due to the model used for the simulations, as will be described in Chapter 8. Furthermore, these measurements with large uncertainties motivate the need for a luminosity upgrade at JLab to measure the charge naturality of J/ψ photoproduction and test the theoretical assumptions on the dominant J/ψ production mechanism, especially given that the calculation of the naturality requires measuring an additional three polarised SDMEs.

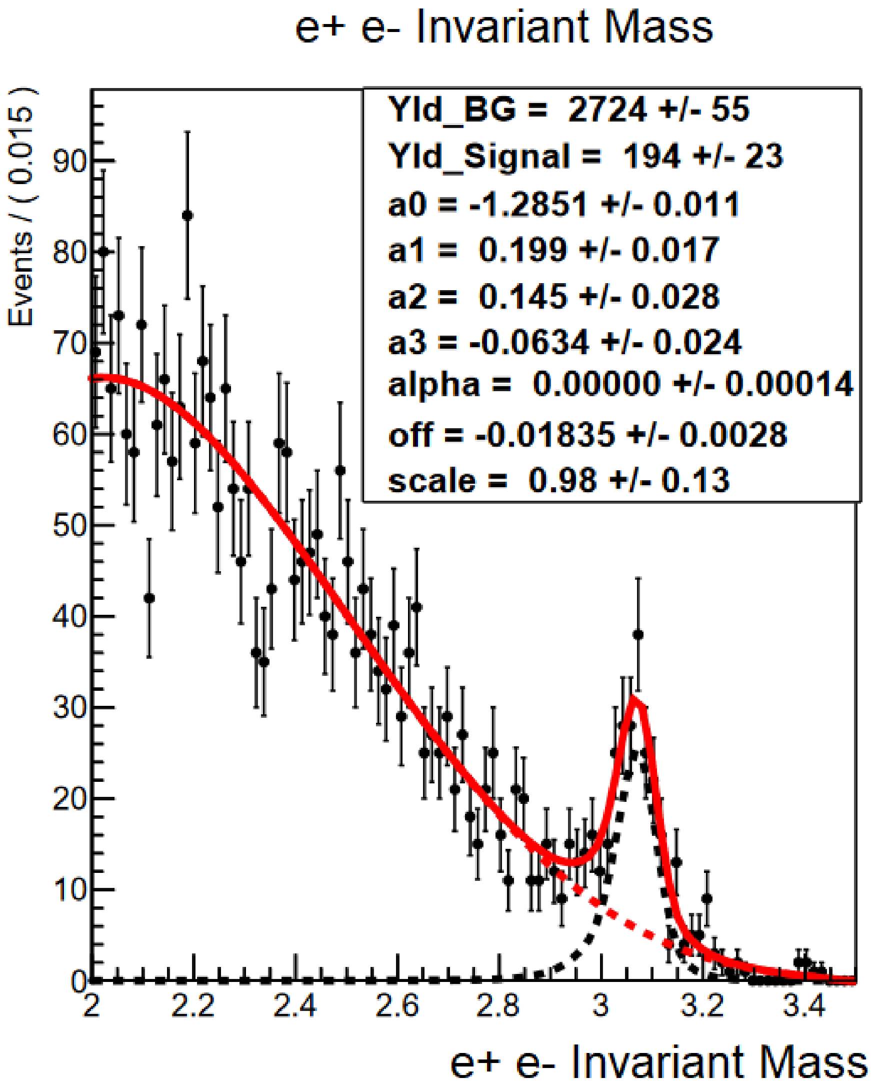


Figure 7.24: The di-lepton invariant mass distribution in the channel $ep_{bound} \rightarrow (e')e^{+}e^{-}p$ fitted with *brufit* and sPlot. The parameters a0 to a3 are the parameters of a third order Chebychev polynomial fitted to the background. The parameters alpha, off, and scale refer to the convolution between the fitted data and the simulated model representing the J/ψ signal which was passed to *brufit*.

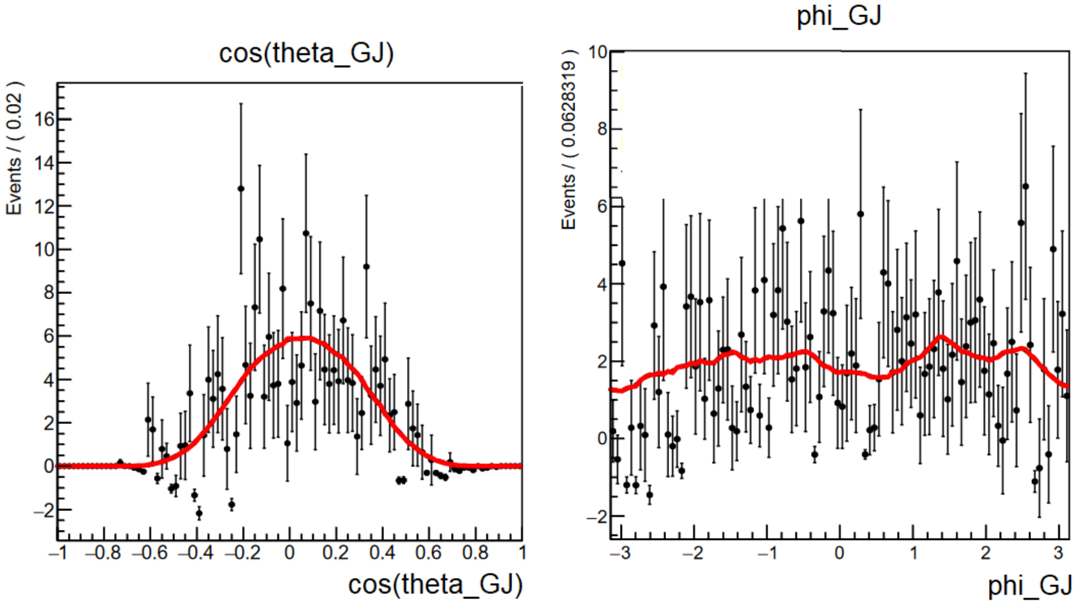


Figure 7.25: The decay angles $\cos(\theta_{GJ})$ (Left) and ϕ_{GJ} (right) fitted using Equation 7.2.

Chapter 8

The Total and Differential Cross Sections

The two previous chapters both focused on analysis techniques aimed at identifying final state particles and events composed of these particles from which the J/ψ photoproduction total and differential cross section can be measured. These measure the probability that J/ψ is produced when the electron beam interacts with the hydrogen or deuteron target. The total cross section $\sigma_0(E_\gamma)$ is measured as a function of quasi-real photon energy. The dependency in E_γ is obtained by measuring the cross section in bins of E_γ. The differential cross section $\frac{d\sigma(E_\gamma)}{dt}$ is measured as a function of t and E_γ, which is obtained by measuring the cross section in bins of E_γ and t. The total and differential cross sections are calculated below using Equations 8.1 and 8.2 respectively.

$$\sigma_0(E_\gamma) = \frac{N_{J/\psi}(E_\gamma)}{N_\gamma \cdot \rho_T \cdot l_T \cdot \omega_c \cdot Br \cdot \varepsilon(E_\gamma)} \tag{8.1}$$

$$\frac{d\sigma(E_\gamma)}{dt} = \frac{N_{J/\psi}(t, E_\gamma)}{N_\gamma \cdot \rho_T \cdot l_T \cdot \omega_c \cdot Br \cdot \varepsilon(t, E_\gamma) \cdot \Delta t} \tag{8.2}$$

$N_{J/\psi}$ refers to the number of J/ψ measured in each bin of quasi-real photon energy or t. This is measured as shown in Chapter 7 by fitting the di-lepton invariant mass and extracting the J/ψ yield from the fit in each E_γ or t bin. As for every detector, CLAS12's detection efficiency is not perfect, with some particles travelling through CLAS12 going undetected. The reconstruction algorithms at CLAS12 are also imperfect, meaning that some particles will not be reconstructed well and will not be available for analysis. This can happen for example when the track finding algorithm for the drift chambers fails and a track is not identified by the overall reconstruction algorithm. Furthermore, CLAS12 has a geometrical acceptance which is related to the detector's physical size and geometry. To state the obvious, a particle that misses the detector cannot be detected. The overall acceptance takes into account the detection and reconstruction efficiency and the geometrical acceptance. From now the term acceptance refers to the overall acceptance. The total amount of J/ψ produced by the interaction of the electron beam with the target is recovered by dividing the number of J/ψ detected by the acceptance ε measured in bins of E_γ

or t. Details of the acceptance calculation are given in the first section of this chapter.

The product $N_\gamma \cdot \rho_T \cdot l_T \cdot Br$ in the denominator of Equations 8.1 and 8.2 normalises the number of J/ψ produced based on the experiment and reaction to give the cross section. N_γ refers to the number of real or virtual photons produced by the interaction of the electron beam with the target. Details on how this is calculated will be given in the second section of this chapter. l_T is the target length, with ρ_T the number of protons or neutrons in the target, with both these terms given in Chapter 4. Br is the branching ratio, which is roughly 6% for the J/ψ decay to a di-lepton pair. The Δt term in the differential cross section calculation is taken as the width of the t bin. The product of N_γ, l_T and ρ_T describe the quasi-real photoproduction luminosity of the experiment. The branching ratio then accounts for the fact that only 6% of J/ψ will decay to an electron positron pair or a di-muon pair.

The final term of Equations 8.1 and 8.2 is an overall normalisation factor ω_c. This corrects for errors in the flux or acceptance calculations which are not necessarily perfect. Details on how this normalisation factor is calculated are given in the third section of this chapter. The penultimate section of this chapter then shows the total and differential cross sections as measured for J/ψ photoproduction on the proton, with J/ψ decaying to a di-muon pair and J/ψ photoproduction on the bound proton and bound neutron, with J/ψ decaying to an electron positron pair. The final section of this chapter then details some studies of the systematic uncertainty in these measurements.

8.1 Acceptance Calculation

The principle behind the acceptance calculation is rather straightforward. As mentioned in Chapter 4, simulated events are first generated then passed to the simulation framework for CLAS12 named *GEMC*. These events are then reconstructed from the simulated hits in the detector using the normal reconstruction algorithms. If the simulation is well matched to CLAS12 data, it will then provide the same detection and reconstruction efficiency and geometrical acceptance as the real detector. The acceptance calculation is then realised by taking the ratio of the true generated events to the reconstructed events in simulated data. Figure 8.1 shows the quasi-real photon energy of the reconstructed and true generated events for the simulation of the $ep_{bound} \rightarrow (e')e^+e^-p$ channel. The ratio of both plots then gives the acceptance as a function of quasi-real photon energy for the $ep_{bound} \rightarrow (e')e^+e^-p$ channel.

Figure 8.2 shows the acceptance as a function of the quasi-real photon energy of the $ep \rightarrow (e')\mu^+\mu^-p$ and $en_{bound} \rightarrow (e')e^+e^-n$ channels. Neutrons have a lower detection efficiency than

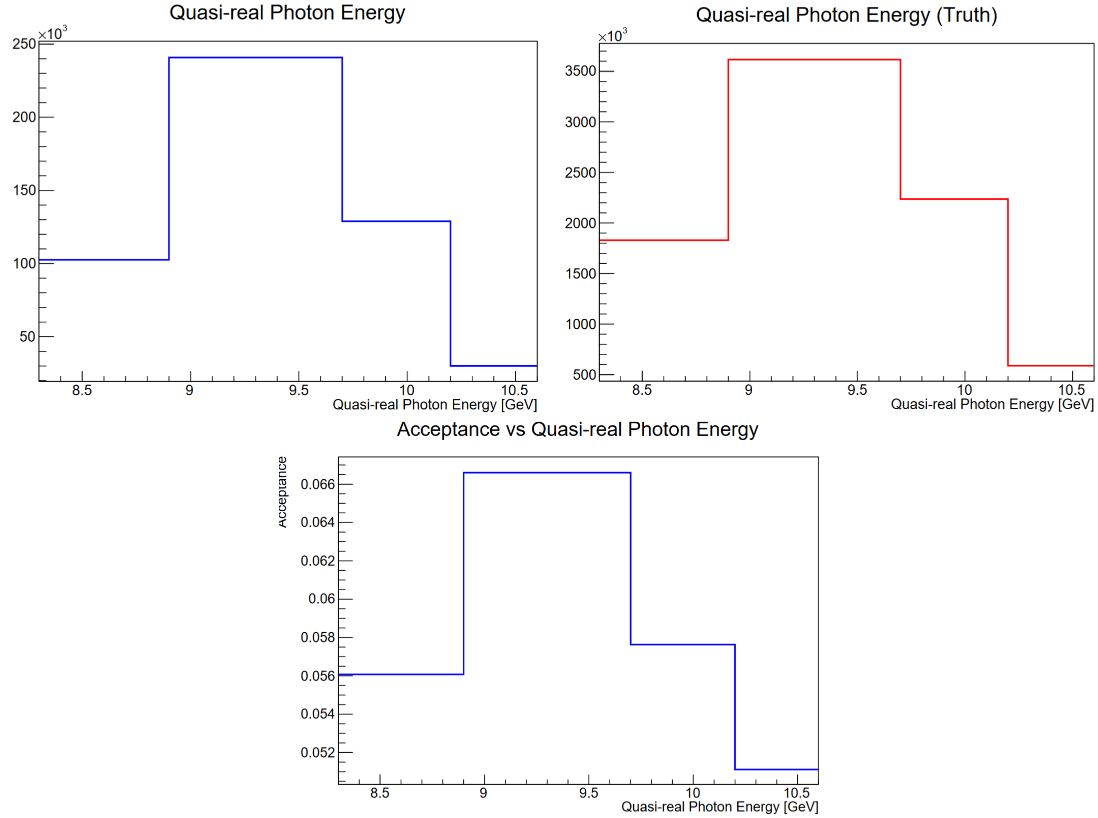


Figure 8.1: Top Left: The quasi-real photon energy of the simulated, reconstructed events. Top Right: The quasi-real photon energy of the true generated simulated events. Bottom: The resulting acceptance as a function of the quasi-real photon energy. All plots are shown for the $ep_{bound} \to (e')e^+e^-p$ channel.

protons, which results in a lower acceptance. The detection efficiency of muons and similarly charged electrons or positrons are roughly equivalent.

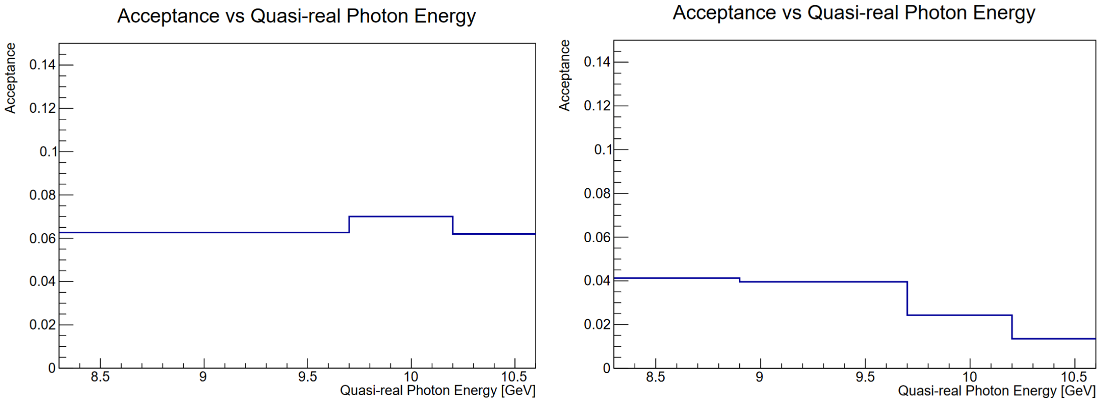


Figure 8.2: The acceptance as a function of the quasi-real photon energy of the $ep \to (e')\mu^+\mu^-p$ (Left) and $en_{bound} \to (e')e^+e^-n$ (Right) channels.

As shown in Chapter 4 the data taken on both a deuteron or hydrogen target used various different beam currents. A higher beam current typically leads to a lower reconstruction efficiency due to a higher occupancy, notably in the drift chambers. The background merging, as detailed in Chapter 4 accounts for the varying background at different beam currents. An example of the acceptance with and without background merging is shown in Figure 8.3 for the $ep_{bound} \rightarrow (e')e^+e^-p$ channel. The simulation was made assuming a 10.6 GeV beam and with roughly half of the data using a 35 nA background and the other half using a 50 nA background to reflect the proportions of CLAS12 data taken at that energy and beam currents, as explained in Chapter 4. The background merging decreases the acceptance by roughly 50%, as it simulates the inefficiencies in the reconstruction algorithm. A drop of 50% given three detected charged particles corresponds to a drop in efficiency of about 80% per charged particle. To make sure that the simulated data is representative of the varying beam current in the CLAS12 datasets, a similar proportion of events are generated with the background merging at a given beam current to the proportion of accumulated charge of the data taken at a given beam current. Similarly, the data taken on a deuteron target used two different beam energies. The simulated datasets on a deuteron target contain a similar proportion of generated events to the proportion of accumulated charge of the data taken at a given beam energy.

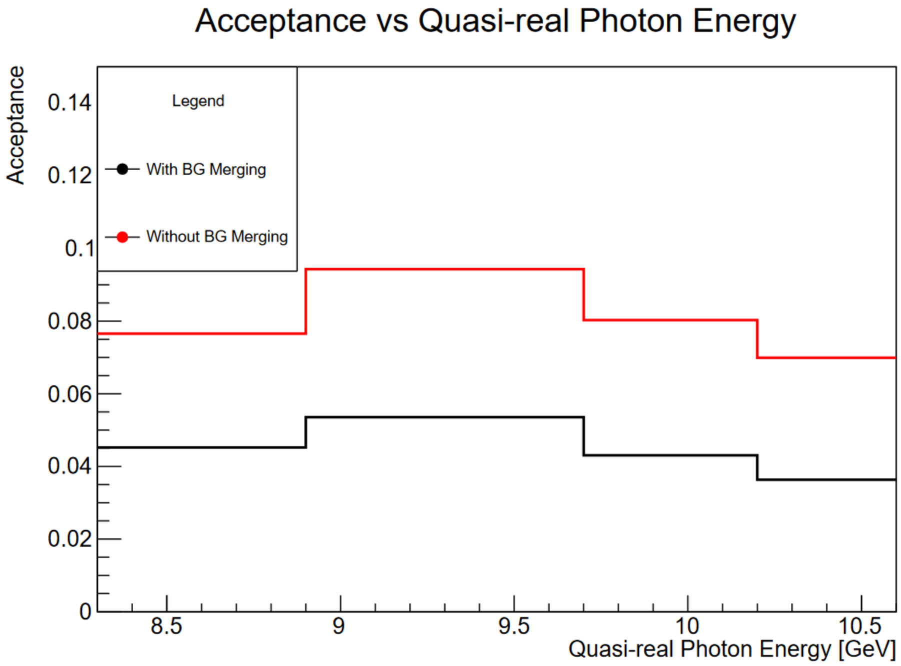


Figure 8.3: The acceptance as a function of the quasi-real photon energy for the $ep_{bound} \rightarrow (e')e^+e^-p$ channel with (Black) and without (Red) background merging.

It is however worth noting that the acceptance plots shown in Figures 8.1 and 8.2 are really just there as examples. Changing the binning scheme will obviously affect these, as would changing the cuts on the missing mass squared or Q^2. For example, Figure 8.4 shows how the acceptance varies when tightening or loosening the cuts on the missing mass squared and Q^2. As expected, tighter cuts lead to a lower acceptance and to fewer measured J/ψ. Overall, assuming that the simulated events are well representative of CLAS12 data, the number of J/ψ and the acceptance will balance each other out to provide a stable cross section regardless of the cut placed on quantities such as the missing mass squared or Q^2.

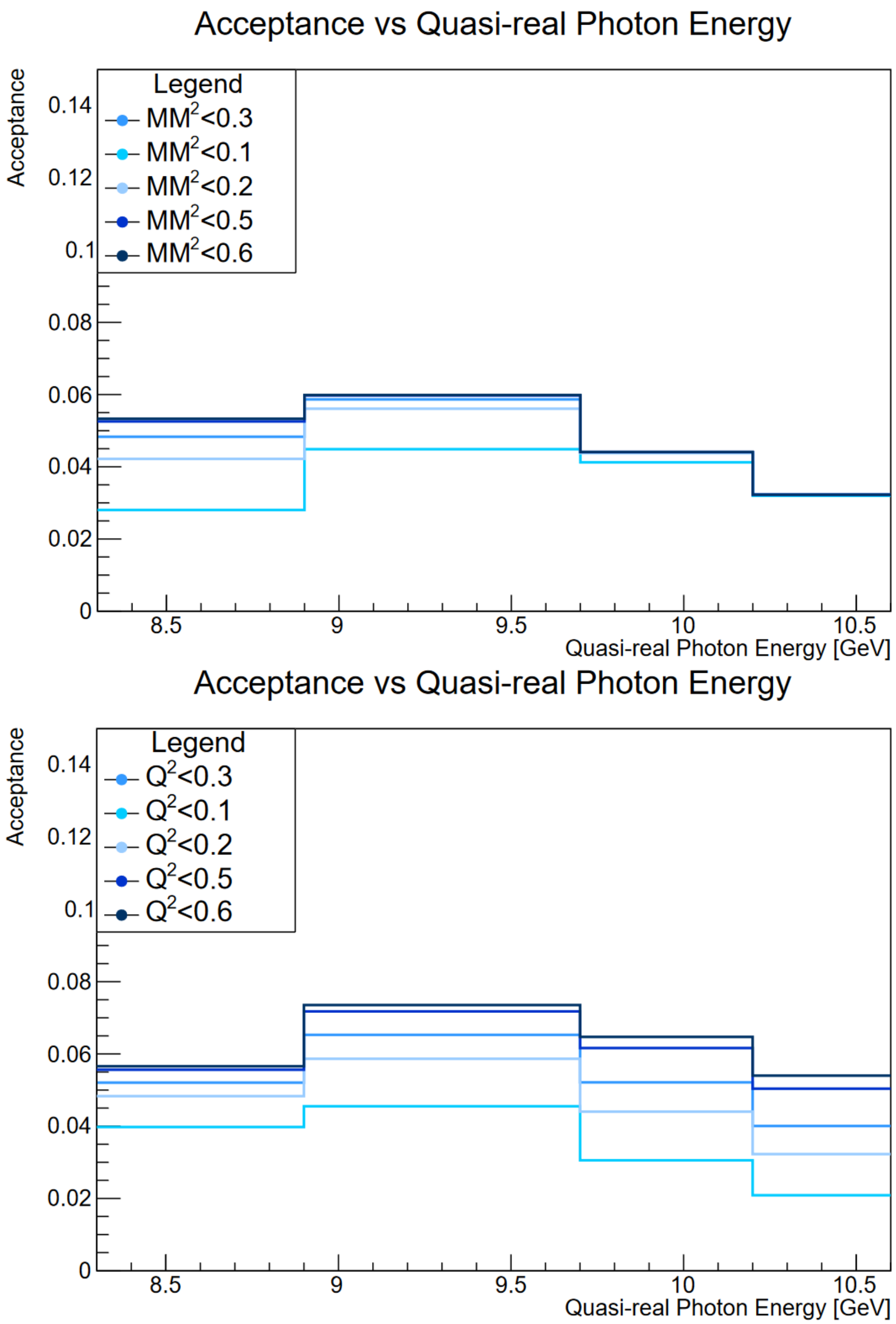


Figure 8.4: The acceptance as a function of the quasi-real photon energy for the $ep_{bound} \rightarrow (e')e^+e^-p$ channel, when varying the cuts on the missing mass squared (top) and Q^2 (bottom).

Machine learning techniques can also be used to map the acceptance as a function of the final state particles' four-momentum. As shown in Appendix A.3, these machine learning algorithms can be used to create a fast simulation accurate within a percent with regards to the conventional simulation but providing a significant speed-up in event rate. However, they could also be used to assign weights to events in CLAS12 data which would allow for a multi-dimensional acceptance correction instead of a one, or at best two, dimensional acceptance correction as shown above. Some work is also ongoing in using CLAS12 data with machine learning techniques to extract the acceptance correction directly from data instead of relying on simulations. Single particle acceptance can be mapped from exclusive channel such as the $ep \to e'n\pi^+$ channel used for the neutron corrections in Chapter 6. This is complicated by non-negligible resolutions on the reconstructed four-momentum and ways to mitigate this are being investigated. Another complication is that at present it is unclear how to extract multi-particle correlations that affect the acceptance of the full event on top of the per particle acceptance. These multi-particle correlations happen for example due to overlapping tracks and are only present when considering two or more particles. As such it is currently hard to estimate the acceptance corrections directly from data but the techniques described above offer a promising direction for acceptance corrections that is being investigated.

8.2 Flux Calculation

As mentioned at the start of this chapter, the number of photons N_γ produced in the interaction of the electron beam with the target forms part of the luminosity and cross section calculation. The number of photons per GeV is defined as:

$$N_\gamma = \frac{Q(F_R + F_V)}{q} \tag{8.3}$$

where Q is the accumulated charge of the dataset, q is the electron charge $q = 1.6 \cdot 10^{-19}C$ and F_R and F_V the real and virtual photon flux respectively. The real photon flux is due to real bremsstrahlung photons produced inside the target and target cell as the electrons from the beam interact with the electric field of the constituent protons and neutrons of the target or target cell. The real photon flux per GeV is calculated as:

$$F_R = \frac{1}{E_\gamma}\left(\frac{l_T}{2X_0\_T} + \frac{l_c}{X_0\_c}\right)\left(\frac{4}{3} - \frac{4}{3} \times \frac{E_\gamma}{E_b} + \frac{E_\gamma^2}{E_b^2}\right) \tag{8.4}$$

for X_0 and l the scattering length and length of the target (T) and target cell (c) and E_b the electron beam energy. The factor of a half for the target contribution comes from the fact that bremsstrahlung photons are produced throughout the target, with a photon produced at the end of the cell having a lower probability of interacting with the target and therefore a lower contribution to the luminosity. This effect works out to a factor of a half on average.

The virtual photon flux is due to the virtual photons mediating the interaction of the electron beam with the target. This can be calculated based on the equivalent photon approximation as:

$$F_V = \frac{1}{E_b} \times \frac{\alpha}{x\pi}((1-x+\frac{x^2}{2})log(\frac{Q^2_{max}}{Q^2_{min}}) - (1-x)) \tag{8.5}$$

where $x = \frac{E_\gamma}{E_b}$, $Q^2_{min} = M_e^2 \frac{x^2}{(1-x)}$ and Q^2_{max} a free parameter, for M_e the mass of an electron and α the fine structure constant. This calculation was then refined by Frixione *et. al.* [150]. Figure 8.5 shows the real photon flux, the EPA virtual photon flux and the Frixione virtual photon flux, where Q^2_{max} is set to 0.02 for both virtual flux calculations. The Frixione flux was here also re-scaled so that it did not exactly overlap with the EPA flux in the plot but the two formulas give the exact same results.

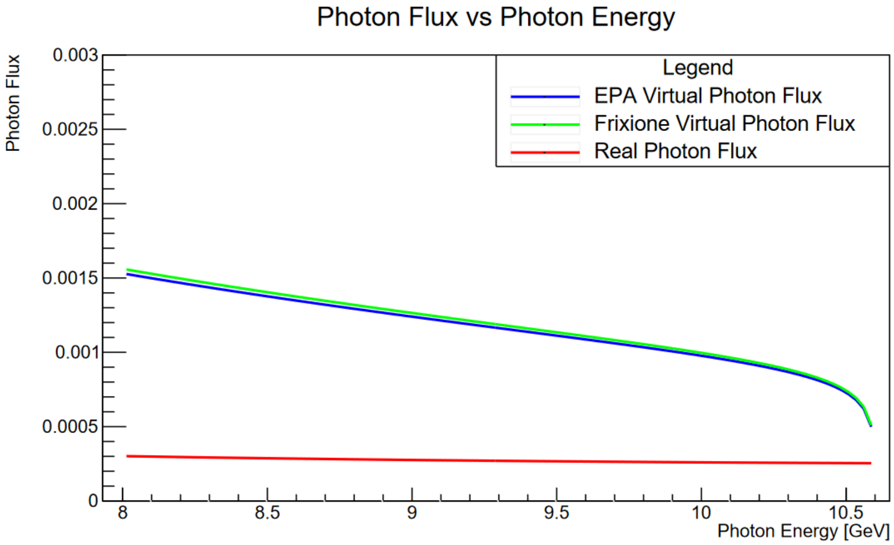


Figure 8.5: The real (red), EPA virtual (blue) and Frixione virtual (green) photon fluxes where Q^2_{max} is set to 0.02 for both virtual flux calculations. The Frixione flux was here also re-scaled so that it did not exactly overlap with the EPA flux in the plot but the two formulas give the exact same results.

As mentioned in the previous paragraph, Q^2_{max} was set to 0.02 in the flux calculations shown in Figure 8.5. This corresponds to the integration limit in Q^2 over which the virtual photon flux is calculated. It can however be extended to $Q^2_{max} = 2E_\gamma M_p$ [150], for M_p the mass of a proton. Similarly, near-threshold effects can affect the Frixione virtual flux calculation at large Q^2 [150]. This is shown in Figure 8.6. The net effect of increasing Q^2_{max} is to increase the virtual photon flux. The near-threshold effects only occur at high Q^2_{max} and decrease the photon flux when the photon energy goes close to threshold (8.2 GeV for J/ψ). The choice of photon flux calculation is in principle accounted for by the normalisation, as explained in the next section, and should not introduce a systematic effect in the cross section calculation.

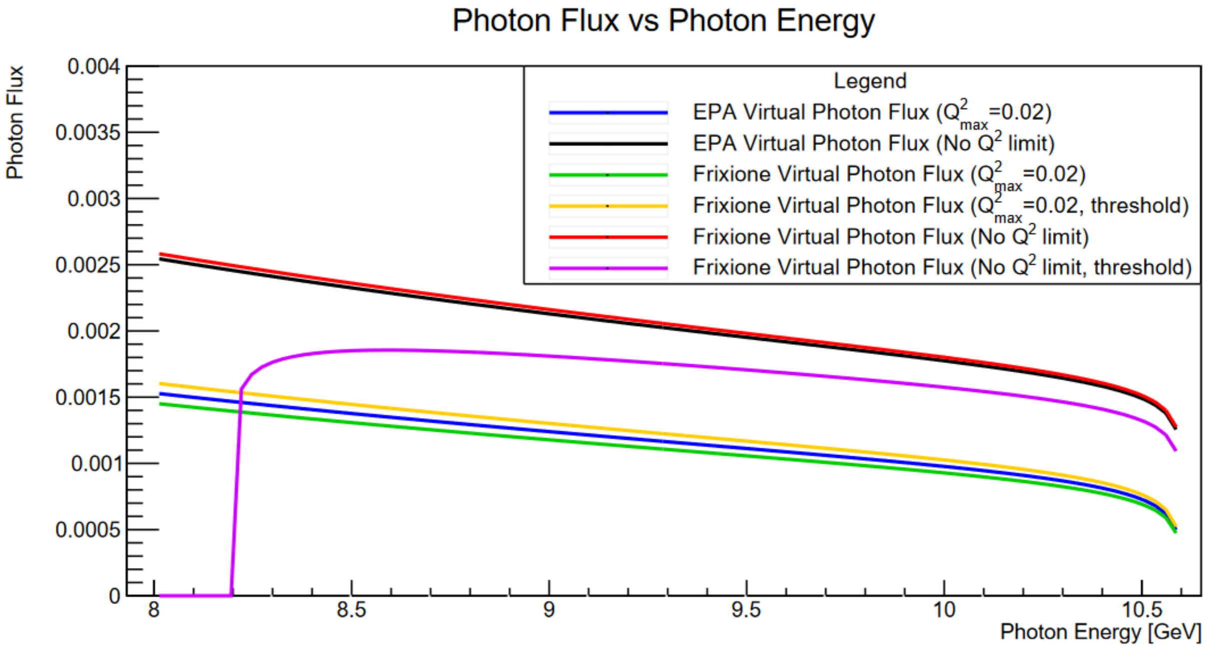


Figure 8.6: The EPA virtual flux with Q^2_{max} set to 0.02 (blue) or $Q^2_{max} = 2E_\gamma M_p$ (black). The Frixione virtual photon flux with Q^2_{max} set to 0.02 (green) and with near-threshold effects accounted for (orange) or with $Q^2_{max} = 2E_\gamma M_p$ (red) and with near-threshold effects accounted for (magenta). Note that nearly overlapping lines would be overlapping if not for a small scaling introduced when plotting.

The real and virtual photon flux calculations as shown in Figure 8.5 are then integrated over E_γ in each E_γ bin, and summed to give the number of photons as described at the start of this section. When several datasets are used to measure the cross sections and these were taken with different beam energy, the sum $Q(F_R + F_V)$ is turned into $\Sigma_i Q_i(F_{R\_i} + F_{V\_i})$ where i denotes the different datasets.

8.3 Normalisation

The next step in the cross section calculation is to calculate an overall normalisation factor. This corrects for the acceptance and flux calculations, giving some flexibility in allowing these not to be perfect. Most of the work on the normalisation has been done by CLAS12 collaborators [146]. As such, an overview is given here to follow the rest of the discussion without going into too much detail.

The strategy for normalisation is based on the well known Bethe Heitler cross section. Bethe Heitler events are simulated as usual and re-weighted using the known cross section and the same flux and luminosity calculation as described previously. As such, the simulation ties in the same flux and acceptance used to calculate the cross section. By calculating the ratio of simulated events to the CLAS12 events in the di-lepton invariant mass range of 2 GeV to 2.9

GeV, a normalisation is established relating the CLAS12 cross section calculation to the known Bethe Heitler cross section [146].

Figure 8.7 shows the simulated electro and photoproduction Bethe Heitler events and the CLAS12 data in the $ep \to (e')e^+e^-p$ channel. The Bethe Heitler simulation is weighted using the cross section and the real or virtual flux depending on whether the simulation is for electro or photoproduction of Bethe Heitler events. A J/ψ simulation is also overlayed, this is weighted by eye to give a comparison to CLAS12 data in the region of the J/ψ mass. The CLAS12 data is first shown in the Blue dotted histogram using similar analysis techniques as to what was described previously. However, an additional correction is established in Black and Red. This is done by considering the final states e^-e^-p and e^+e^+p. These do not conserve charge and should in principle only contain background events produced in a similar fashion to the background in the $ep \to (e')e^+e^-p$ channel. The same charge lepton final states then allow to produce a correction factor that can be applied to the $ep \to (e')e^+e^-p$ channel. Once this correction is applied, the normalisation factor is taken from the ratio of the simulated and weighted Bethe Heitler events to the CLAS12 events [146].

Here the normalisation factor was extracted with the EPA virtual flux with $Q^2_{Max} = 0.02$. Table 8.1 shows the ω_c extracted as described above in each E_γ bin used for the analysis of J/ψ produced on hydrogen target, decaying to an electron positron pair.

E_γ Range	ω_c
8.4-8.9 GeV	0.35
8.9-9.3 GeV	0.43
9.3-9.7 GeV	0.34
9.7-10.2 GeV	0.48
10.2-10.6 GeV	0.34

Table 8.1: The ω_c normalisation in each E_γ bin of the J/ψ analysis in the $ep \to (e')e^+e^-p$ channel.

As mentioned above, the normalisation was only extracted for a final state with an electron positron pair and a proton, for the dataset taken on a hydrogen target. One might then question applying it to different datasets and different final state particles. The first justification for this is that the main source of inefficiencies comes from drift chamber tracking. These should be similar for muons and electrons. Another observation is that the normalisation is roughly equal whether the proton is in the central or forward detector [146], suggesting that the main sources of inefficiencies might come from tracking highly relativistic particles such as electrons or muons in the forward detector. As such it seems as if the normalisation can be applied to final states with a recoil proton or neutron. It must however be noted that the normalisation is

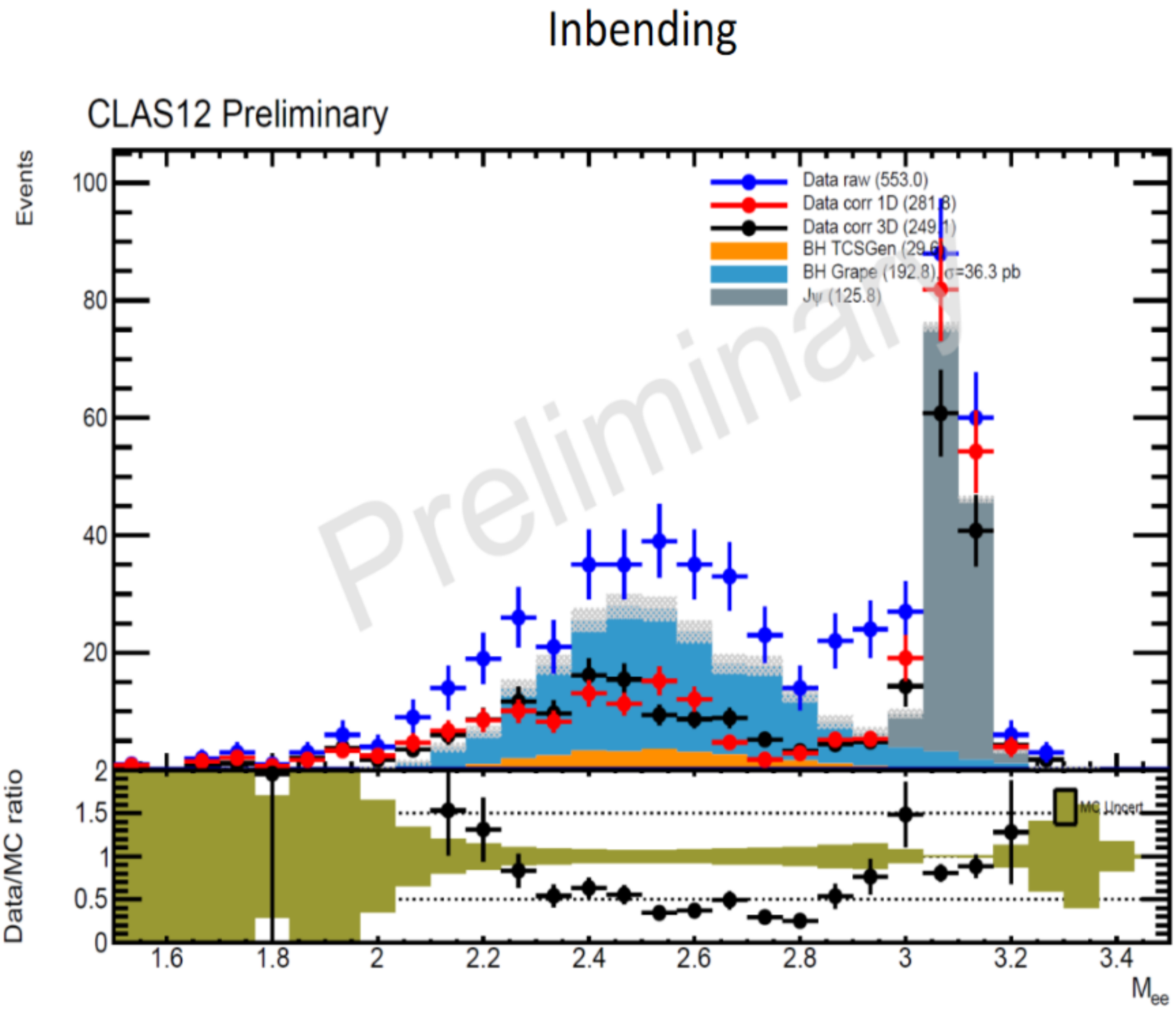


Figure 8.7: The di-lepton invariant mass for $ep \to e^+e^-p$ events. The simulated data is represented by the filled histograms, with a J/ψ simulation (Grey) and Bethe Heitler photoproduction (Orange) and electroproduction (Blue) simulations. CLAS12 data is represented by the dotted histograms, before applying the same charged lepton correction (Blue) and with the 1D (Red) and 3D (Black) corrections. A ratio of CLAS12 data to simulation is shown underneath the invariant mass spectra. Taken from [146].

very large, and it may be that the low statistics in the analysis used to calculate the normalisation hide small differences between final states. Finally, although the drift chamber occupancy and therefore tracking inefficiency will change for a dataset taken on a hydrogen target to one taken on a deuteron target, in principle the background merging should account for differences between datasets. The overall normalisation and background correction process is still being refined [146]. As such, for now, there is no point in simply repeating this analysis whilst it is still incomplete. The normalisation established above will therefore be applied to all channels investigated in this thesis as a first order approximation to what the actual normalisation should be. The systematic uncertainty that the normalisation introduces on the extraction of the proton and neutron mass radii will be calculated in the last section of this chapter.

The effect of the normalisation on the J/ψ total cross section is shown in Figure 8.8 for the $ep_{bound} \rightarrow (e')e^+e^-p$ channel. As the binning scheme is different to what is shown in Table 8.1, an average normalisation of $\omega_c = 0.388$ was used here. As shown the effect is significant, driving the cross section much closer to the J/ψ total cross section as measured at GlueX. From now on this average normalisation of $\omega_c = 0.388$ will be used for all cross section calculations, in conjunction with the EPA virtual flux with $Q^2_{max} = 0.02$.

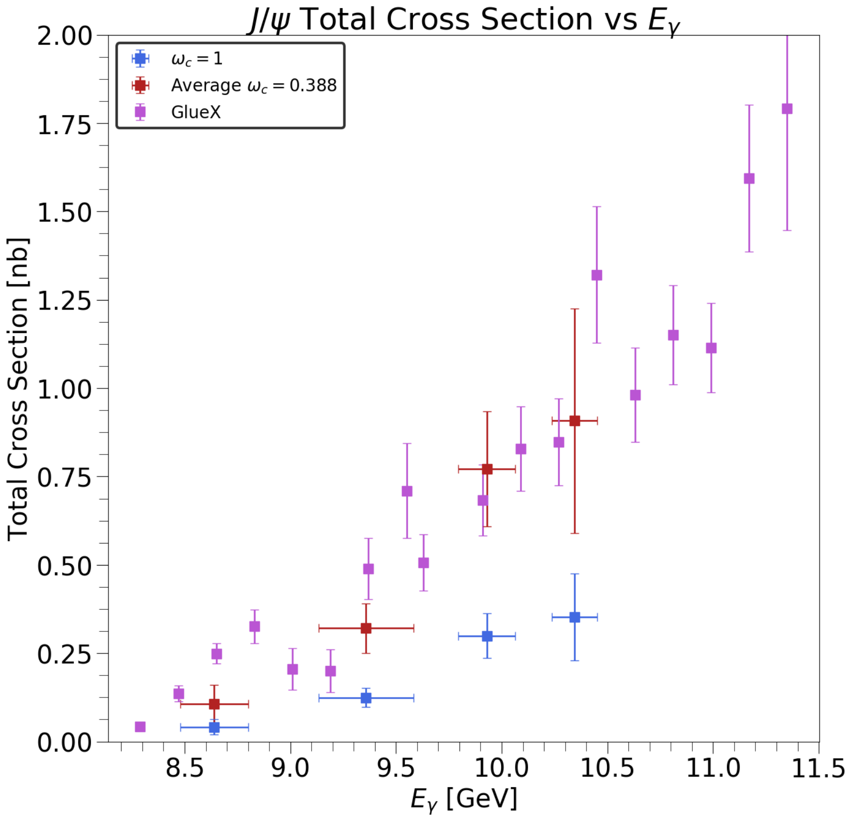


Figure 8.8: The J/ψ total cross section, measured in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel with $\omega_c = 1$ (blue) or $\omega_c = 0.388$ (red) and compared to the GlueX results (magenta).

8.4 The Total and Differential Cross Sections

The previous sections have detailed how all the necessary ingredients for the cross section can be calculated. In fact, Figure 8.8 showed the J/ψ total cross measured in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel and compared this with the most recent GlueX results. This section will present first the J/ψ total cross section measured in the $ep \rightarrow (e')\mu^+\mu^-p$, $ep_{bound} \rightarrow (e')e^+e^-p$ and $en_{bound} \rightarrow (e')e^+e^-n$ channels compared to each other, to the GlueX measurements and to the J/ψ total cross section measured in the $ep \rightarrow (e')e^+e^-p$ channel as measured by CLAS12 collaborators [146]. Note that the results measured in this thesis are plotted at the mean of the E_γ or t distribution within the given bins, with the uncertainty in E_γ or t quantifying the error on the mean calculation. This is not the case for the results from GlueX or those provided by collaborators for J/ψ produced on the free proton and decaying to an electron positron pair. Next, this section will present the results in these same channels for the differential cross section and the proton, bound proton and bound neutron mass radii. The discussion of the importance and validity of these results is left to Chapter 9. The invariant mass plots for each bin used in the total or differential cross section measurements and for each channel discussed here are shown in Appendix B.

The first result to note is the J/ψ total cross section measured in the $ep \to (e')\mu^+\mu^- p$ channel, as shown in Figure 8.9. This constitutes the first measurement at CLAS12 requiring muons in the final state. It also presents independent measurements that are in good agreement with those made in the $ep \to (e')e^+e^- p$ channel and with the GlueX measurements.

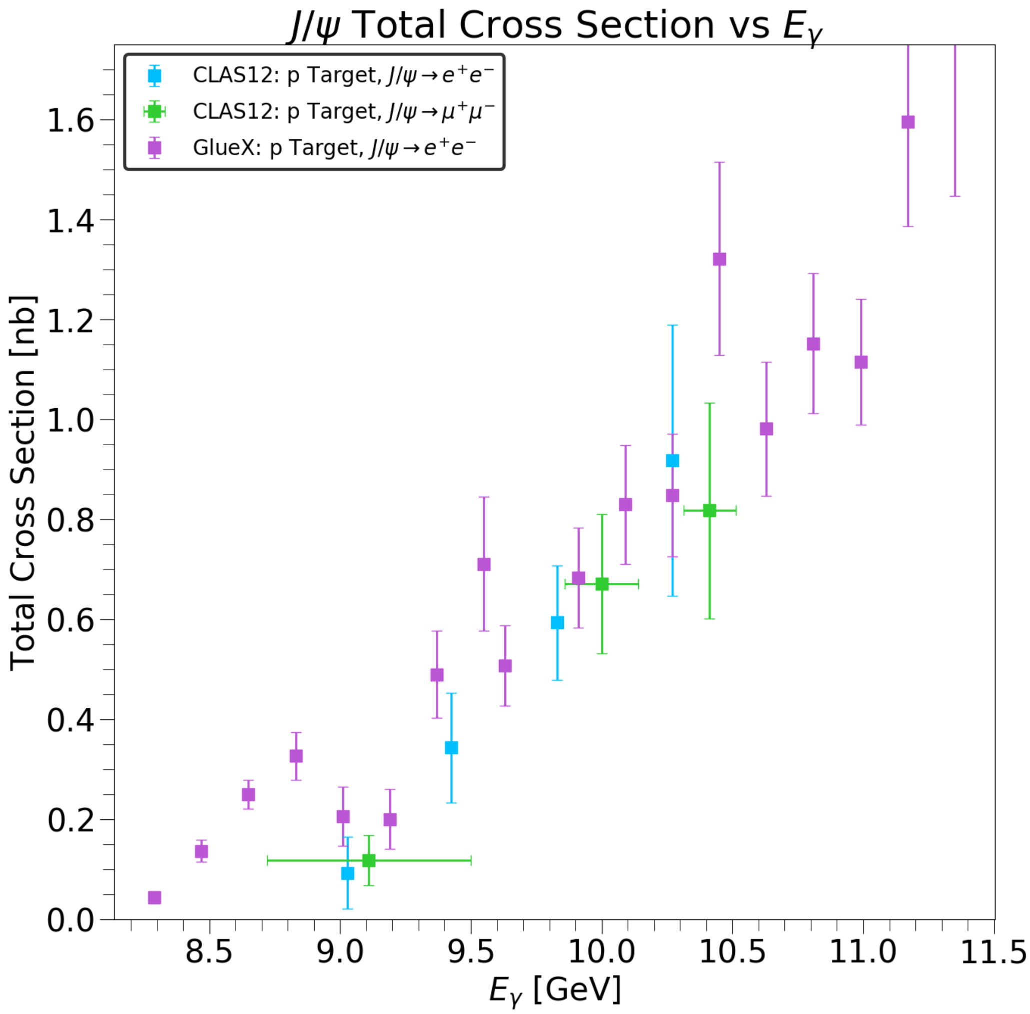


Figure 8.9: The J/ψ total cross section, measured in the $ep \to (e')\mu^+\mu^- p$ channel (green) and $ep \to (e')e^+e^- p$ channel (cyan) compared to the GlueX results (magenta).

Figure 8.10 shows a comparison of the J/ψ total cross section measured on the free or bound proton. The cross section off the free and bound proton agree within the statistical uncertainties, suggesting that any contributions due to final-state interactions in the quasi-free sample are smaller than the statistical precision of the measurement. This constitutes a first measurement of the J/ψ total cross section on the bound proton in the near-threshold region.

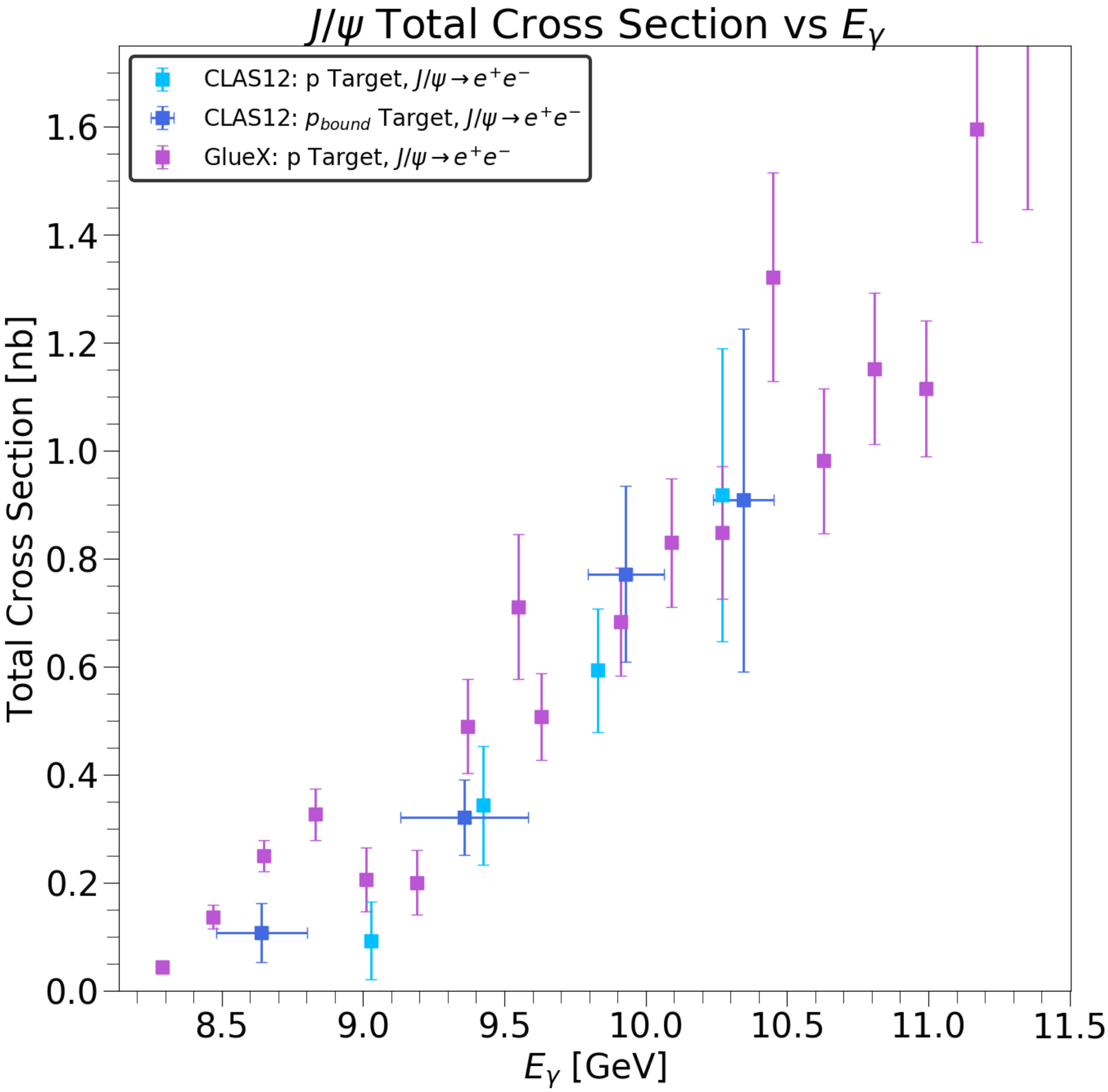


Figure 8.10: The J/ψ total cross section, measured in the $ep_{bound} \to (e')e^+e^-p$ channel (blue) and $ep \to (e')e^+e^-p$ channel (cyan) compared to the GlueX results (magenta).

Figure 8.11 shows a comparison of the ϕ total cross section measured on the free or bound proton. Both cross sections agree well with each other. Although a slight departure from the J/ψ measurements, the interest in measuring the ϕ total cross section is twofold. First, this allows to test the assumption that the normalisation is valid when a proton is *not* detected in the forward detector, as the proton mainly goes in the central detector for the ϕ cross section measurement. Agreement was found with the preliminary analysis from another experiment [151]. This gives weight to the assumption that the normalisation is also valid when replacing the forward detector proton with a forward detector neutron in the J/ψ cross section measurement. Secondly, measuring the ϕ cross section can be related to the scattering length calculation, as mentioned in Chapter 2. Further studies of the ϕ production mechanism may support or not the validity of extracting the mass radius of the nucleon from ϕ photoproduction.

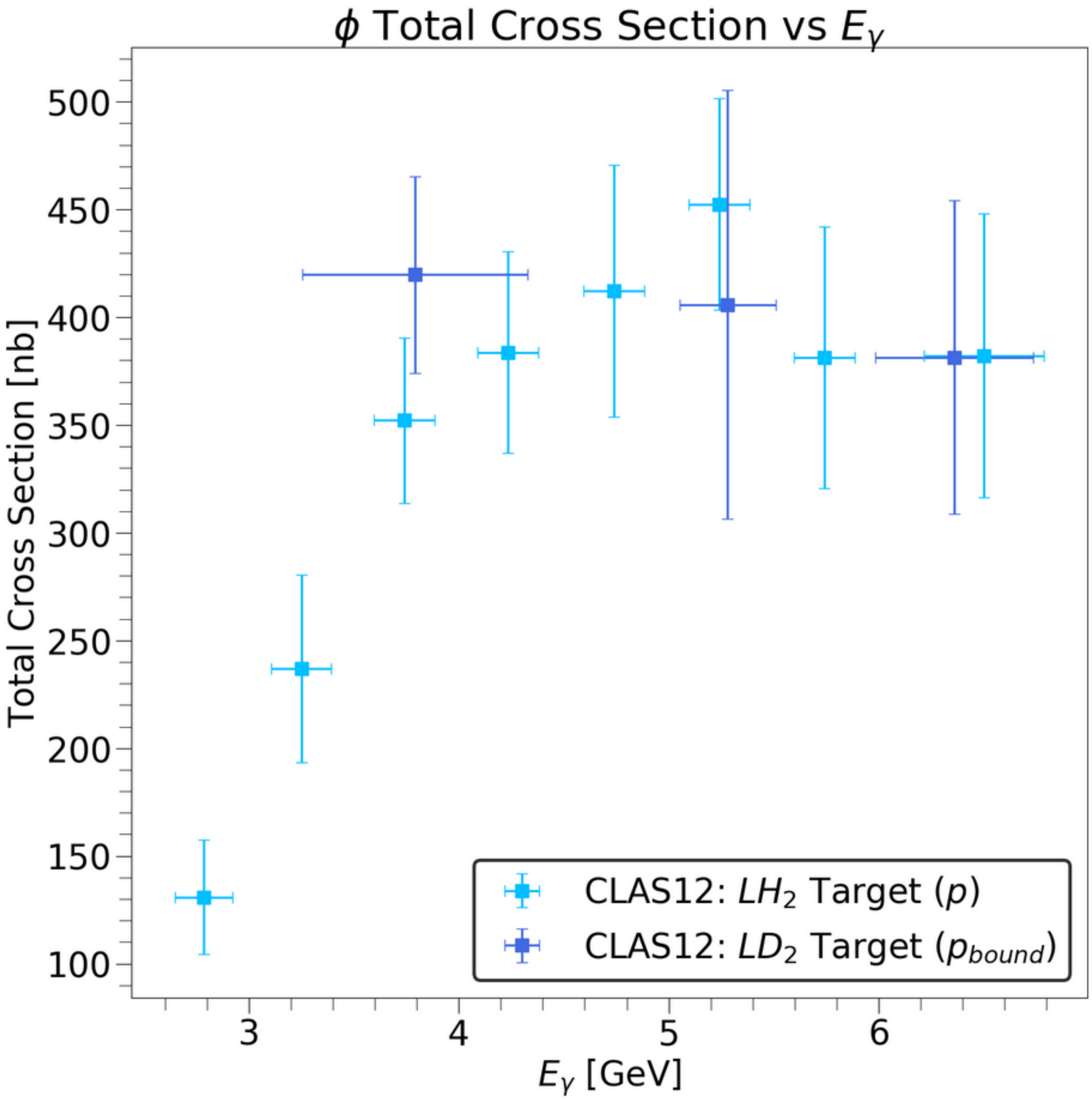


Figure 8.11: The ϕ total cross section, measured in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel (blue) and $ep \rightarrow (e')e^+e^-p$ channel (cyan).

Figure 8.12 shows a comparison of the J/ψ total cross section measured on the bound proton or bound neutron, and compared to the GlueX measurements on the free proton. This constitutes a first measurement of the J/ψ total cross section on the neutron in the near-threshold region. The agreement between the cross sections on the proton and neutron also supports the assumption of a two- or three-gluon exchange as the dominant production mechanism, as these should be isospin invariant.

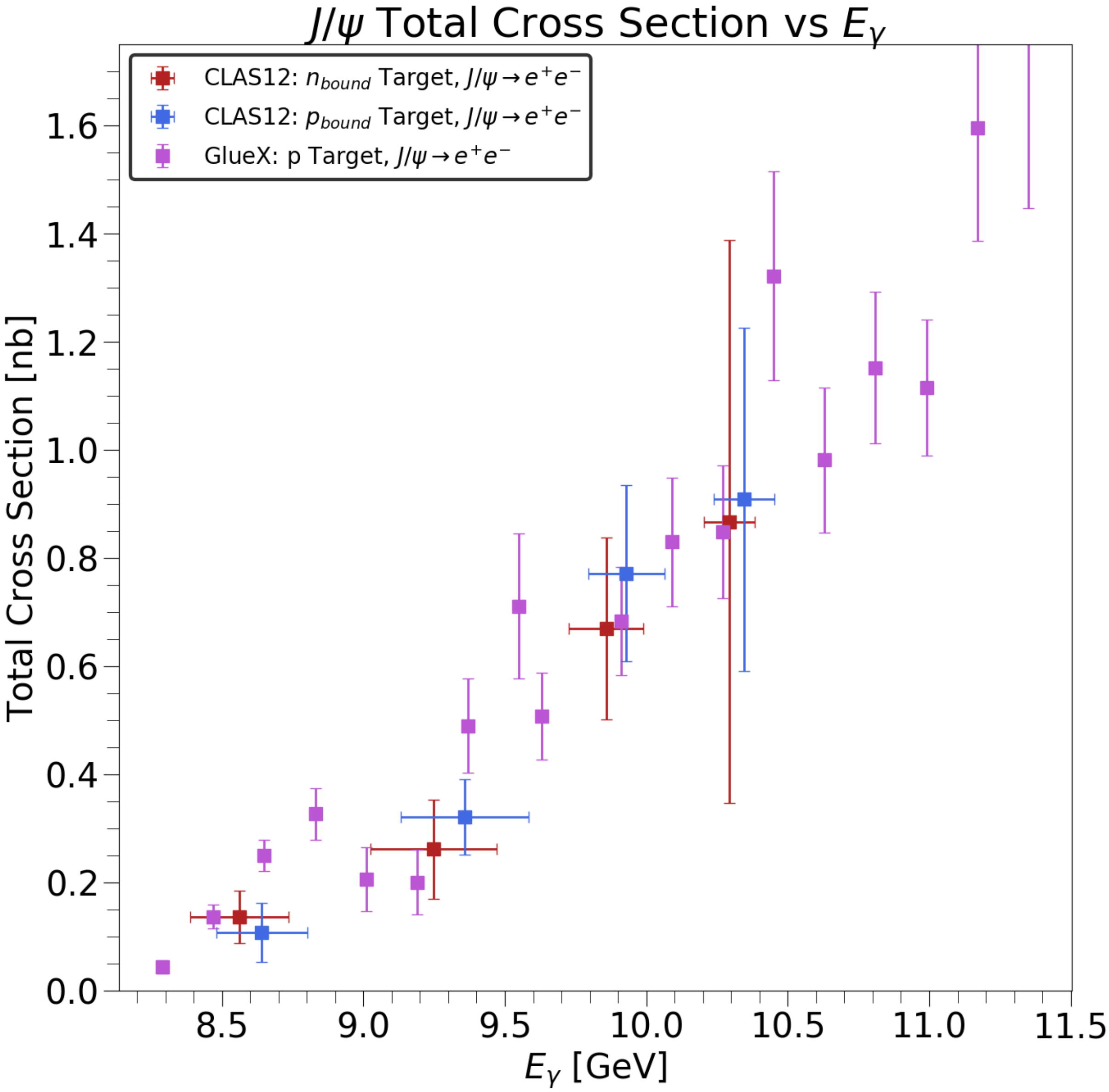


Figure 8.12: The J/ψ total cross section, measured in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel (blue) and $en_{bound} \rightarrow (e')e^+e^-n$ channel (red) compared to the GlueX results (magenta).

Figure 8.13 shows a comparison of the J/ψ total cross section measured on the bound proton when using or not the machine learning based reaction identification tool developed and presented in Chapter 7. As shown, especially in the third bin, even a low cut on the response of 0.15 has a significant impact on the total cross section measurement. The largest divergence occurs at a response cut of 0.86 which corroborates the conclusions established from Figures 7.12 and 7.13. Overall it should be rather clear by now that these types of tools are unsuited to cross section measurements as they introduce a bias that can increase the total cross section.

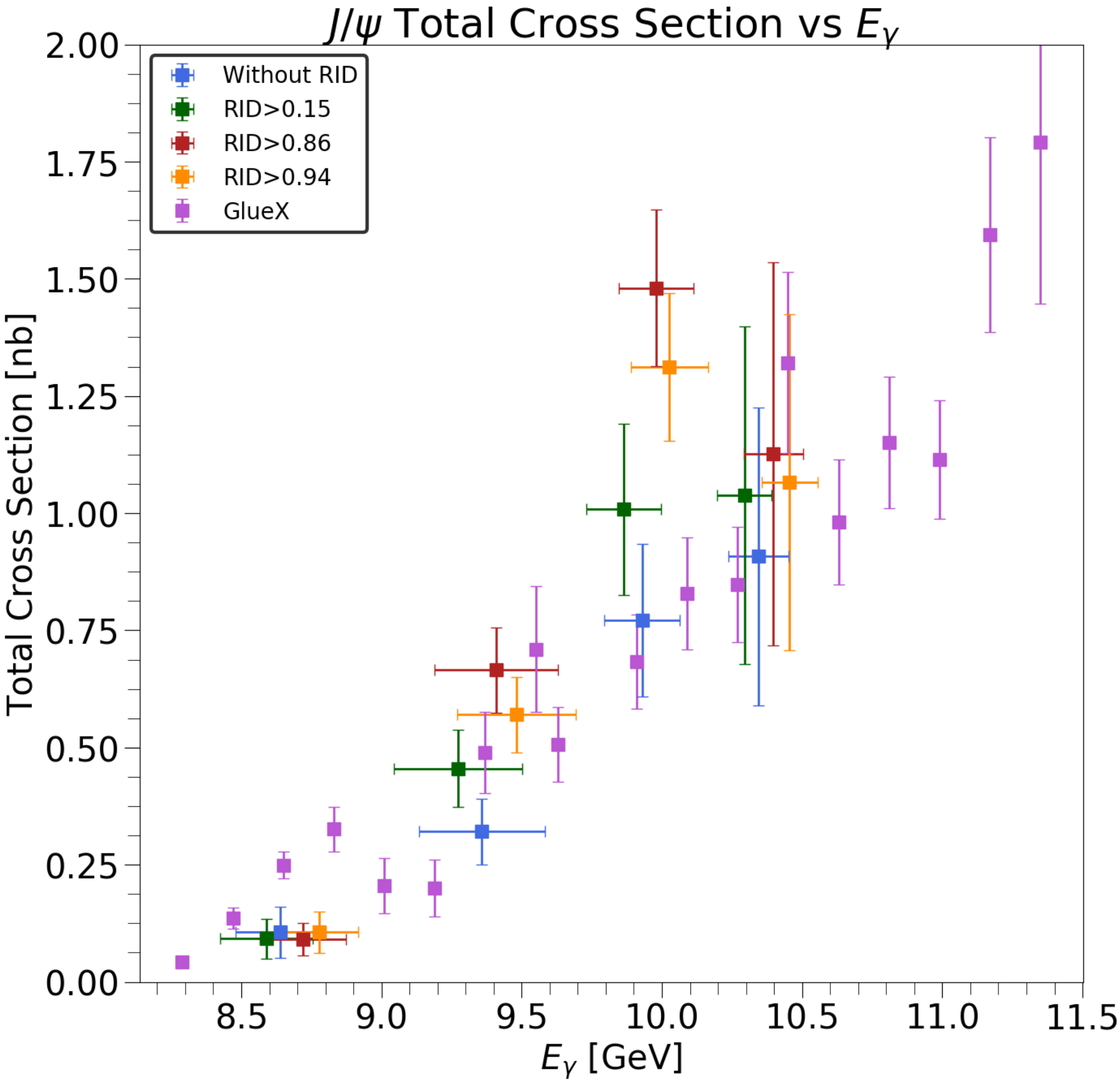


Figure 8.13: The J/ψ total cross section, measured in the $ep_{bound} \to (e')e^+e^-p$ channel with the machine learning based reaction ID and compared to the GlueX results (magenta). The cut on the response is set at 0.15 (green), 0.86 (red) and 0.94 (orange), or the reaction ID tool is not used (blue).

As explained in Chapter 2, the nucleon mass radius can be measured by fitting the differential cross section as a function of t. Specifically, Equation 2.9 relates the differential cross section to a gravitational form factor which can be expressed with a dipole parametrisation as shown in Equation 2.11. This is changed slightly to include a scale factor p_s which takes into account the differential cross section at $t = 0$ and the nucleon mass squared, such that the fitted function $f(t)$ is written as:

$$f(t) = p_s(1 - \frac{t}{m_s^2})^{-4} \tag{8.6}$$

The other free parameter m_s and its uncertainty σ_{ms} is measured by fitting the form factor to the differential cross section. The mass radius R_m is then calculated from m_s based on Equation 2.12, with a factor $\hbar c$ converting m_s in GeV to R_m in fm. The uncertainty σ_{Rm} on R_m is given by:

$$\sigma_{Rm} = \hbar c \sqrt{12} m_s^{-2} \sigma_{ms} \tag{8.7}$$

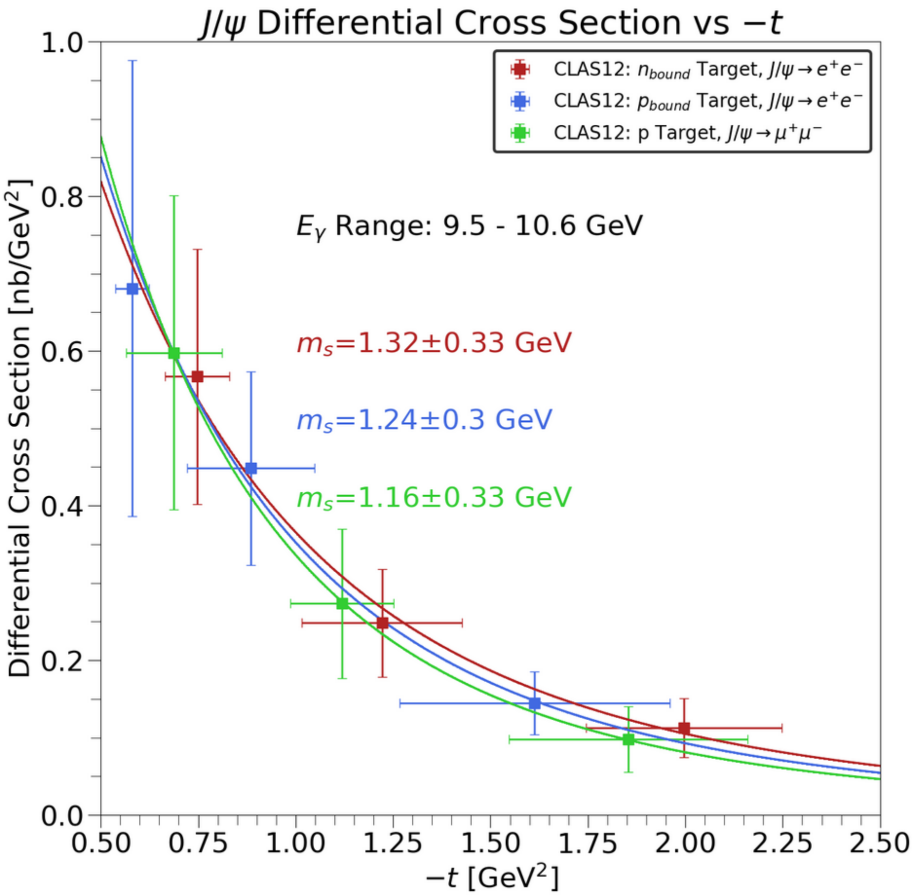


Figure 8.14: The J/ψ differential cross section, measured in the $ep_{bound} \to (e')e^+e^-p$ channel (blue), the $en_{bound} \to (e')e^+e^-n$ channel (red) and the $ep \to (e')\mu^+\mu^-p$ channel (green).

Figure 8.14 shows the differential cross section as a function of $-t$, fitted with the form factor of Equation 2.11 for J/ψ decaying to an electron positron pair, produced on the bound proton and bound neutron, and J/ψ produced on a proton decaying to a di-muon pair. All differential cross sections were measured in an E_γ bin ranging from 9.5 GeV to 10.6 GeV. This range will be reduced when additional statistics are available for the measurement. The scale and shape of the differential cross sections agree well in all three channels.

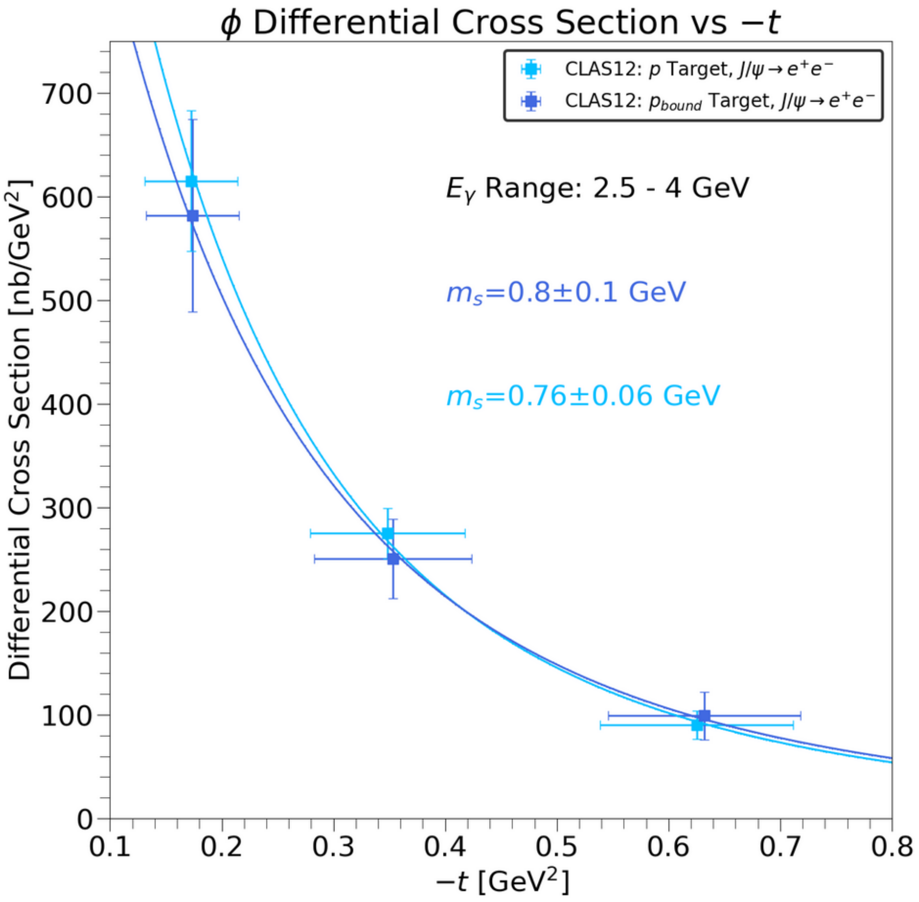


Figure 8.15: The ϕ differential cross section, measured in the $ep \to (e')e^+e^-p$ channel (cyan) and the $ep_{bound} \to (e')e^+e^-p$ channel (blue).

Chapter 2 also described previous measurements of the mass radius of the nucleon based on ϕ photoproduction rather than J/ψ photoproduction. The differential cross section of ϕ decaying to an electron positron pair, produced on the proton and bound proton are shown in Figure 8.15. These were measured in an E_γ bin ranging from 2.5 GeV to 4 GeV. The scale and shape of the differential cross sections agree well in both channels.

The measured m_s and R_m are shown in Table 8.2 and Figure 8.16. This represents a first measurement from J/ψ photoproduction of the mass radius of the bound neutron and bound proton. The mass radius of the proton, bound proton and bound neutron are comparable in size. Good

Experiment	Target	Decay	m_s	R_m
GlueX	p	$J/\psi \to e^+e^-$	1.39±0.09 GeV	0.49±0.03 fm
Hall C	p	$J/\psi \to e^+e^-$	Not reported	0.52±0.03 fm
lQCD	N	N/A	N/A	0.517
CLAS12	p	$J/\psi \to e^+e^-$	1.20±0.30 GeV	0.57±0.14 fm
CLAS12	p	$J/\psi \to \mu^+\mu^-$	1.16±0.33 GeV	0.59±0.17 fm
CLAS12	p_{bound}	$J/\psi \to e^+e^-$	1.24±0.30 GeV	0.55±0.13 fm
CLAS12	n_{bound}	$J/\psi \to e^+e^-$	1.32±0.33 GeV	0.52±0.13 fm
CLAS	p	$\phi \to K^+K^-$	Not Reported	0.67 ± 0.10 fm
CLAS12	p	$\phi \to e^+e^-$	0.85±0.07 GeV	0.81±0.05 fm
CLAS12	p_{bound}	$\phi \to e^+e^-$	0.91±0.11 GeV	0.75±0.08 fm

Table 8.2: The measured m_s and R_m in each channel and compared to previous measurements and lattice QCD (lQCD) predictions for the nucleon.

agreement is also shown when measuring the mass radius of the proton in the di-muon or electron positron decay. All measurements are also consistent with lattice QCD prediction. A larger proton and bound proton mass radii are observed when measuring them with ϕ photoproduction. This agrees with the previous measurements presented in Chapter 2 and will be discussed in more detail in Chapter 9. Note that the previous measurements of the mass radius from ϕ photoproduction are quoted in the E_γ bin closest to that used for the CLAS12 measurements. The measurements from GlueX are averaged over the two bins covering the E_γ range used in this thesis, whilst the results from Hall C are averaged for E_γ above 9.7 GeV and up to 10.6 GeV.

Table 8.3 shows the fitted scale factor p_s for all three J/ψ photoproduction channels studied in this thesis. The fitted p_s terms present very large statistical uncertainties which makes it impossible to derive any meaningful information from them. The issue here is that the fit to the scale is poorly constrained by only including three bins in the differential cross section. As all three bins are above $-t > 0.5$ GeV^2, the fit is badly extrapolated to $t = 0$ GeV^2, leading to the large uncertainty. It should be noted however that the original description of the gravitational form factor used to fit the differential cross section does not contain this scale factor [57], and the most interesting information is gained from the shape of the cross section encoded in the m_s parameter.

Experiment	Target	Decay	p_s
CLAS12	p	$J/\psi \to \mu^+\mu^-$	3.10±3.17 GeV^2
CLAS12	p_{bound}	$J/\psi \to e^+e^-$	2.64±2.13 GeV^2
CLAS12	n_{bound}	$J/\psi \to e^+e^-$	2.25±1.88 GeV^2

Table 8.3: The fitted p_s in each channel studied in this thesis.

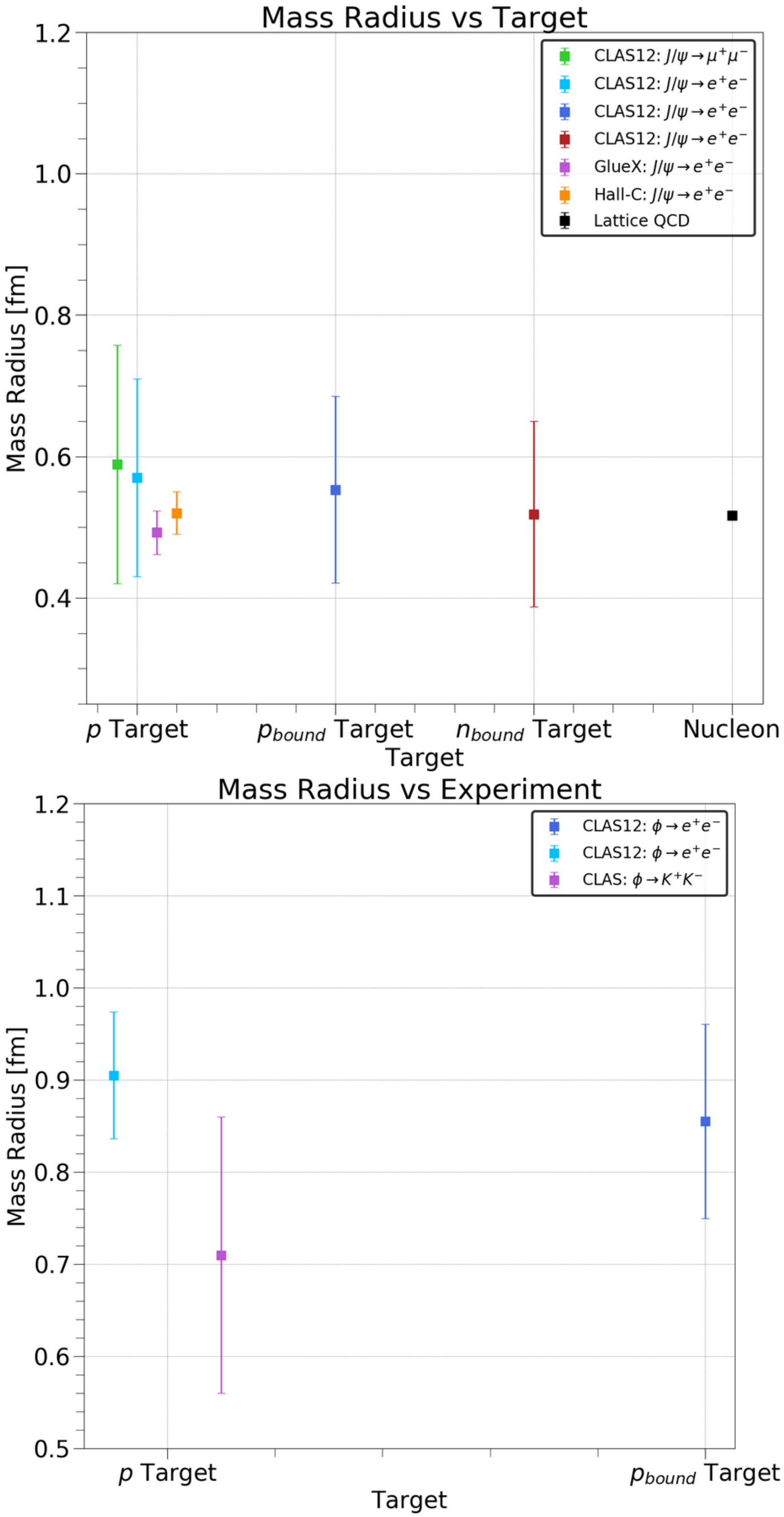


Figure 8.16: The mass radius of various targets measured in J/ψ and ϕ photoproduction for the different experiments and decay channels.

All three J/ψ photoproduction channels were fitted together, and the channels on the bound proton and bound neutron were also fitted together, as shown in Figure 8.17. The first fit can be interpreted as extracting the mass radius of the bound and free nucleon, whilst the second extracts the mass radius of the bound nucleon. Table 8.4 shows the m_s, mass radius and p_s resulting from the combined fits. The statistical uncertainty is reduced by roughly a factor of $\sqrt{(}3)^{-1}$ when combining all channels or the bound proton and neutron channels. Note that the systematic uncertainty associated with using different datasets has not been evaluated when combining all three channels. This should not be an issue when combining the bound proton and neutron channels as the data for these measurements was taken within the same experiment. This then allows to comment on the mass radius of the bound nucleon with a higher statistical precision.

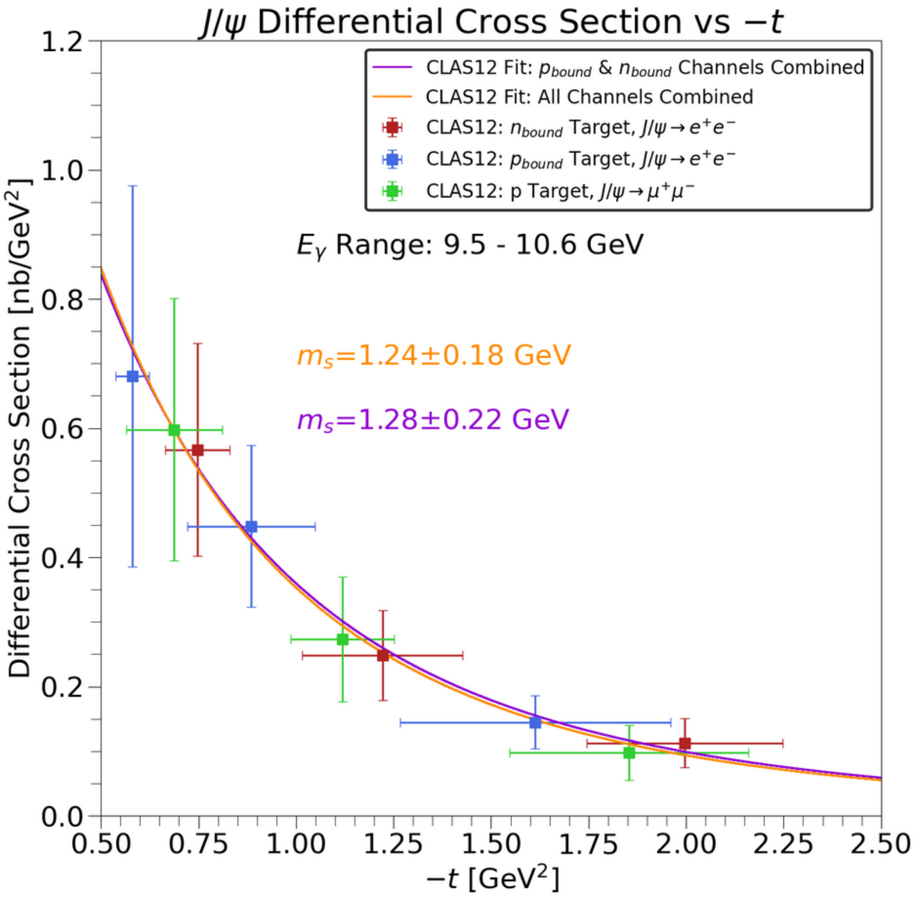


Figure 8.17: The J/ψ differential cross section, measured in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel (blue), the $en_{bound} \rightarrow (e')e^+e^-n$ channel (red) and the $ep \rightarrow (e')\mu^+\mu^-p$ channel (green). A fit is made to all three channels (orange) and to the channels produced on the bound proton and bound neutron (violet).

Combination	m_s	R_m	p_s
All Channels	1.24± 0.18 GeV	0.55±0.08 fm	2.61 ± 1.32 GeV^2
p_{bound} and n_{bound}	1.28±0.22 GeV	0.54±0.09 fm	2.45 ± 1.42 GeV^2

Table 8.4: The measured m_s, R_m and p_s when combining all three channels measured in this thesis or combining the channels on the bound proton and neutron.

8.5 Systematic Uncertainty Quantification

Systematic effects are defined as deviations in results outwith statistical or counting effects due to some aspect of the measurement or analysis. Systematic uncertainties then account for these systematic effects and quantify uncertainties which are not of statistical origin [152]. As the data used for J/ψ photoproduction at CLAS12 was produced independently from this thesis, the focus in quantifying systematic uncertainties will be placed on those due to analysis procedures. As such when evaluating systematic uncertainties introduced in an analysis, the first step should be to estimate whether a deviation in the results is statistically significant when changing some part of the analysis. The Barlow significance S [152] is defined as:

$$S = \frac{\rho_n - \rho_f}{\sqrt{|\sigma_f^2 - \sigma_n^2|}} \tag{8.8}$$

for ρ_f the measured result, such as the cross section, with an optimal set of analysis procedures and ρ_n the result when varying one of the analysis procedures. σ_f and σ_n then account for the statistical uncertainty on the measurements. The Barlow significance then accounts for deviations in results and statistical uncertainty due to a change in the analysis.

The systematic effect due to each analysis procedure can then be quantified by varying each analysis procedure, such as cuts on the missing mass squared or on the response of machine learning classifiers. If the significance is less than two, the change in result is dominated by statistical effects and there are no systematic effect introduced by this procedure [152]. If the significance is greater than four, then there is clear evidence of systematic effects that are not of statistical origin [152]. When the significance is between two and four the decision is left to whoever is performing the analysis to decide whether this denotes a systematic effect [152]. One simple way of thinking about this is that if several different values of a cut all return a significance between two and four then this can denote a systematic effect. If this only occurs for one specific value out of several for a given procedure, then this effect is probably not a systematic effect. In the case where a systematic effect has been identified, it should be corrected. For example by removing the procedure from the analysis. If it is not correctable, then it is necessary to quantify the uncertainty associated with this effect. Equation 8.8 is then modified by introducing a term σ_s in the denominator:

$$S = \frac{\rho_n - \rho_f}{\sqrt{|\sigma_f^2 - \sigma_n^2 - \sigma_s^2|}} \tag{8.9}$$

The magnitude of σ_s is evaluated such that this brings the significance below the threshold of two.

For the analysis of J/ψ photoproduction at CLAS12 the significance of cuts on quantities such as the missing mass squared, Q^2 and the parametrisations of both as a function of E_γ, cuts on the energy deposition of muons, cuts on the distance from the edge of the PCAL for electrons, cuts on the response of the machine learning based positron and muon identification and introducing the JPAC or CLAS12 SDMEs into the simulated models used for the acceptance calculation. The effect of the corrections to the neutron momentum and removal of secondary neutrons are not evaluated, as these are fundamental to the analysis of J/ψ photoproduction on the neutron. Systematic effects introduced by the detectors, for example due to kinematic dependent detection inefficiencies, or introduced by the reconstruction algorithm will not be evaluated here as this is far beyond the scope of this thesis.

The significance of these analysis procedures will be evaluated on the total cross section, as this is expected to affect the differential cross section in much the same way. For clarity, the significance of an analysis procedure repeated across analyses, such as for example cuts on the missing mass squared, are evaluated on the $ep_{bound} \rightarrow (e')e^+e^-p$ channel. This is the channel with the most statistics, and therefore most likely to see any systematic effects, and the cut procedures are shown in previous Chapters to affect different channels similarly.

The overall strategy consists in varying the different cuts from the values of the cut used to calculate the total cross section in the previous chapter. Several sensible values for these cuts are used to evaluate the cross section. The significance is calculated as in Equation 8.8. If the significance falls above four, or between two and four for more than half of the cut values, the cut is taken as having a systematic effect on the analysis. If not, the cut is taken as introducing no systematic effect.

Figure 8.18 shows the significance evaluated for several values of cuts on the missing mass squared and Q^2. Figure 8.19 shows the significance evaluated for several values of cuts on the missing mass squared and Q^2 parametrisations. Varying exclusivity cuts such as those on the missing mass squared or Q^2 also allows to test for any systematic effects due to non-exclusive contributions to the cross section. If no systematic effect is observed when varying these cuts then this is a good further indication that there is no non-exclusive background, or at least that the current measurements do not have the statistical precision to measure the contribution from non-exclusive backgrounds. Figure 8.20 shows the significance evaluated for several values of cuts on the energy deposition of muons in the ECIN, ECOUT and PCAL. These are varied for both muons at the same time. The range of values is limited by the trigger which places upper bounds beyond which no muon will have fired the trigger. Figure 8.21 shows the significance evaluated for several cuts on the response of machine learning classifiers used to identify positrons or di-muon pairs, including when the classifiers are not used. Figure 8.22 shows the significance when varying the cut on the distance of electron clusters from the edge of the PCAL. Finally, the simulated model for J/ψ photoproduction is varied by weighting the simulation used for the acceptance calculation with Equation 7.2 and the measured and predicted SDMEs, as described in Chapter 7. The significance of a more detailed simulation can then be evaluated, as shown in Figure 8.23, to check for systematic effects introduced by a rather simplistic simulated model. Overall, no systematic effect is observed due to any of the cuts or analysis procedures. This is not surprising due to the low statistics available for J/ψ photoproduction at CLAS12, with the analysis dominated by statistical uncertainty. This also motivates not evaluating the systematic effect of each cut on each channel.

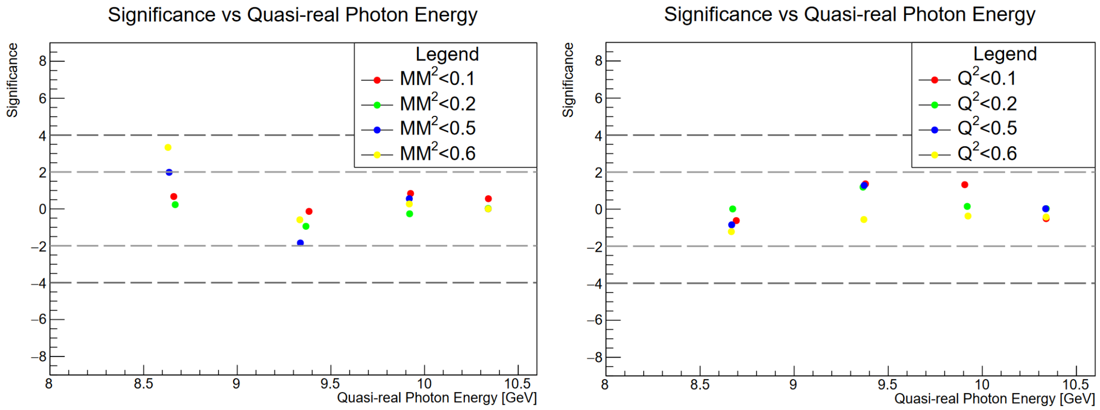


Figure 8.18: The Barlow Significance evaluated when varying cuts on the missing mass squared (left) and Q^2 (right). The optimal cut on the missing mass squared is set as $|MM^2| < 0.3$ GeV^2 and for Q^2 this is set as $Q^2 < 0.3$ GeV^2. Some points may hide each other. No points are hidden by the legend or sit outside of the plot ranges.

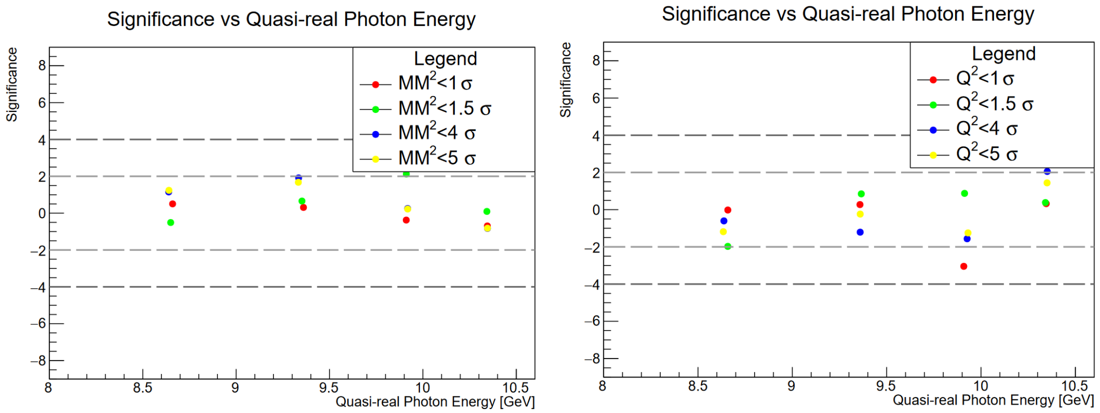


Figure 8.19: The Barlow Significance evaluated when varying cuts on the E_γ parametrisations of the missing mass squared (left) and Q^2 (right). The optimal cut on the missing mass squared is set as $|MM^2| < 2.5\sigma_{MM^2}(E_\gamma)$ and for Q^2 this is set as $Q^2 < 2.5\sigma_{Q^2}(E_\gamma)$. Some points may hide each other. No points are hidden by the legend or sit outside of the plot ranges.

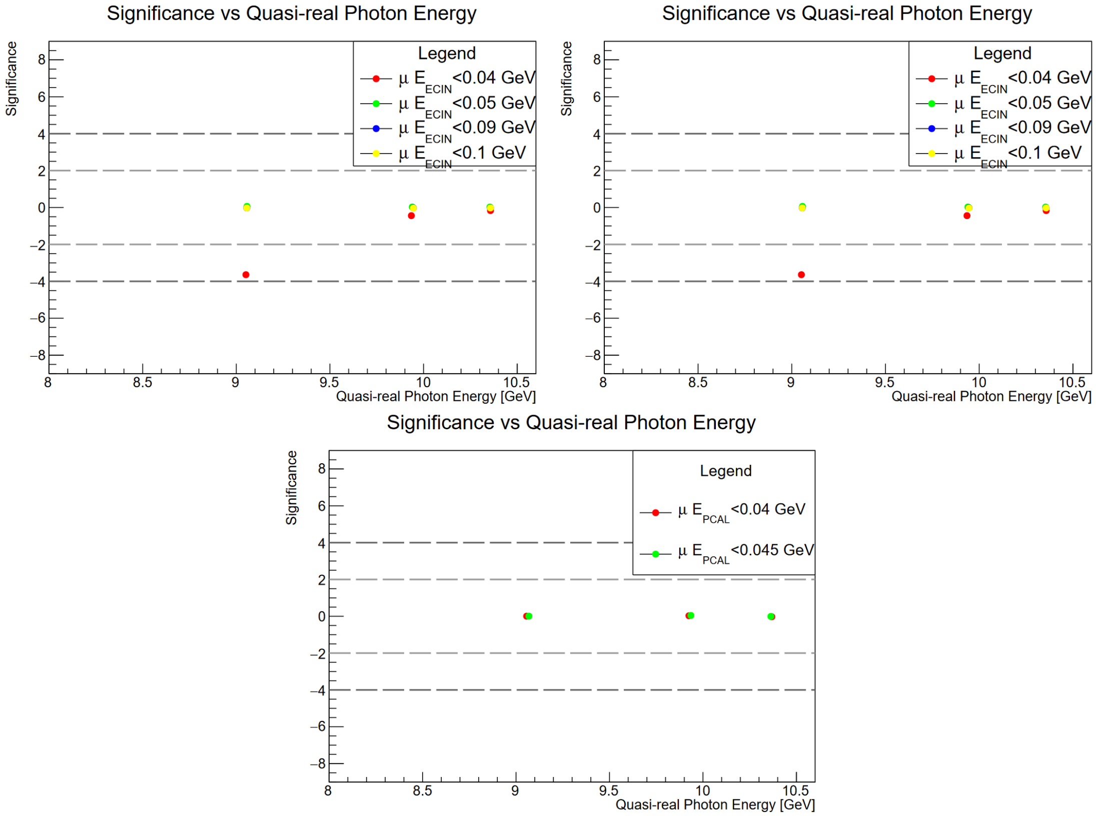


Figure 8.20: The Barlow Significance evaluated when varying cuts on the energy deposition of the di-muon pair in the ECIN (top left), ECOUT (top right) and PCAL (bottom). The cuts are varied for both muons at the same time. The optimal cuts are set as $E_{ECIN} < 0.08$ GeV, $E_{ECOUT} < 0.11$ GeV and $E_{PCAL} < 0.06$ GeV. Some points may hide each other. No points are hidden by the legend or sit outside of the plot ranges.

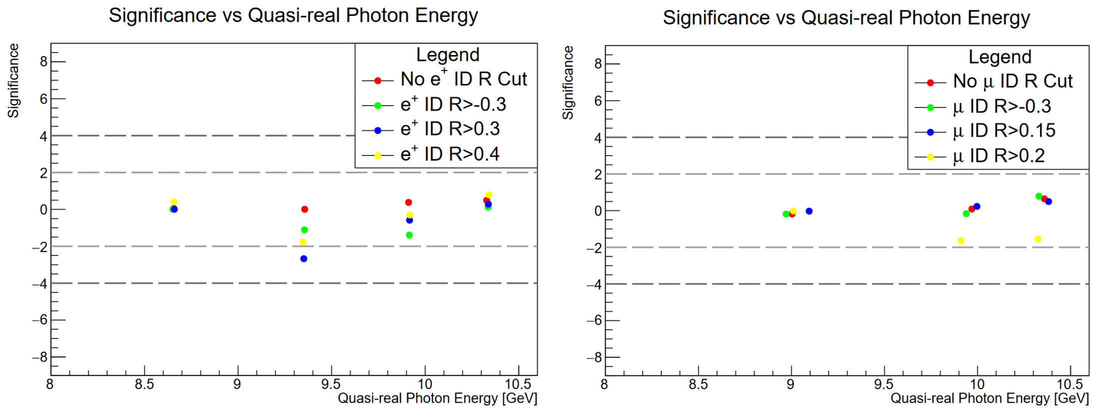


Figure 8.21: The Barlow Significance evaluated when varying cuts on the response of machine learning classifiers used to evaluate a positron (left) and a di-muon pair (right). The optimal cut on the positron identification response is set at e^+ ID R>0.1, the optimal cut on the di-muon identification response is set at μ ID R>0.1. Some points may hide each other. No points are hidden by the legend or sit outside of the plot ranges.

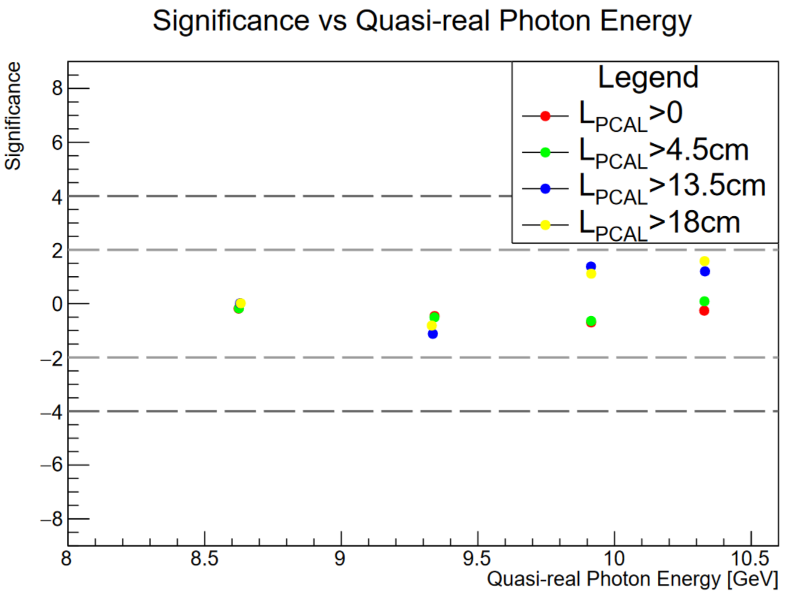


Figure 8.22: The Barlow Significance evaluated when varying cuts on the distance of an electron cluster from the edge of the PCAL. The optimal cut value is set at $L_{PCAL} > 9cm$. Some points may hide each other. No points are hidden by the legend or sit outside of the plot ranges.

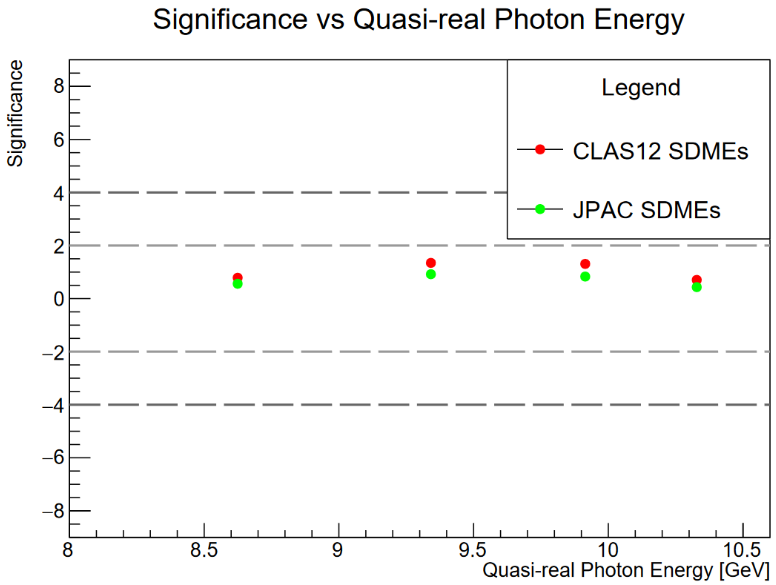


Figure 8.23: The Barlow Significance evaluated when varying the simulated model by weighting the original simulations using Equation 7.2 and the SDMEs measured at CLAS12 or predicted by JPAC. Some points may hide each other. No points are hidden by the legend or sit outside of the plot ranges.

Several analysis procedures for the measurement of the mass radii do not come into play in the total cross section calculation. These will be evaluated for the bound proton, with the assumption that they carry through equally to the bound neutron and free proton. First, a choice of E_γ range in which to calculate the differential cross section might affect the extraction of the mass radii, as the production mechanism of J/ψ may vary as a function of E_γ, as shown in Figure 2.6. The significance is investigated by measuring the mass radius of the bound proton while varying the E_γ range. The upper and lower limits of the E_γ range are varied separately, with Figure 8.24 showing no clear evidence of a systematic effect. It should be noted that the significance is below zero when reducing the range, which points to an overall smaller mass radius, although within statistical uncertainty from the optimal measurement. This is in line with the measurements from Hall C presented in Chapter 2 and could be indicative of a region where the assumption of a two-gluon exchange as the dominant production mechanism is invalid.

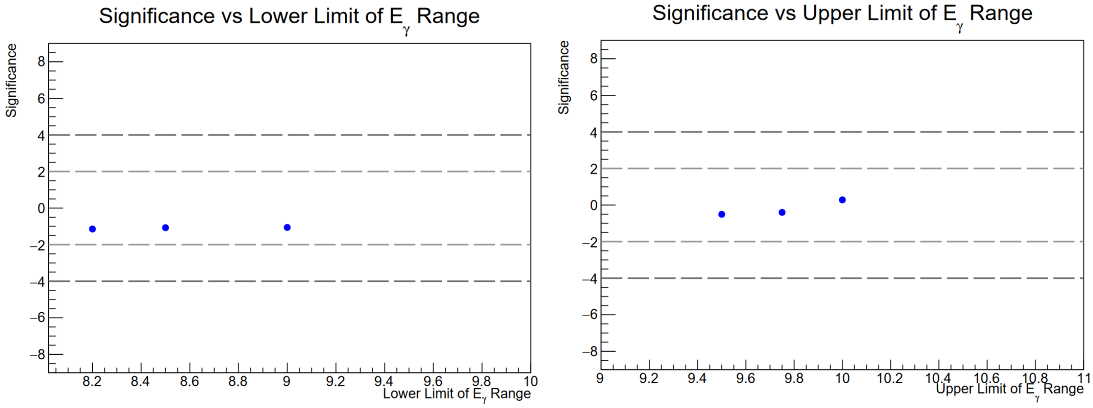


Figure 8.24: The Barlow Significance as a function of the lower (left) or upper (right) limit of the E_γ range used to measure the mass radius of the bound proton. The optimal range is set as $9.5 < E_\gamma < 10.6$ GeV when varying the lower limit and set to $8.0 < E_\gamma < 10.6$ GeV when varying the upper range.

Another important aspect of the analysis is the normalisation factor ω_c. This will obviously have a significant systematic effect on the scale of the total and differential cross sections, as it is a constant factor multiplying the cross sections and scaling them up and down. The systematic uncertainty on the total cross section was not evaluated, as the error on the normalisation is not yet known, and would affect the total cross section calculation. The significance on the measurement of the bound proton mass radius can however be evaluated. As shown in Figure 8.25, the normalisation does not affect the extraction of the mass radius, assuming this normalisation is not t dependent.

Finally, as pointed out in [63], the measured mass radius is dependent on the assumed form factor model. Equation 2.11 assumes a dipole parametrisation, but a monopole, tripole or quadrupole model could also be conceivable. Figure 8.26 shows the significance as a function

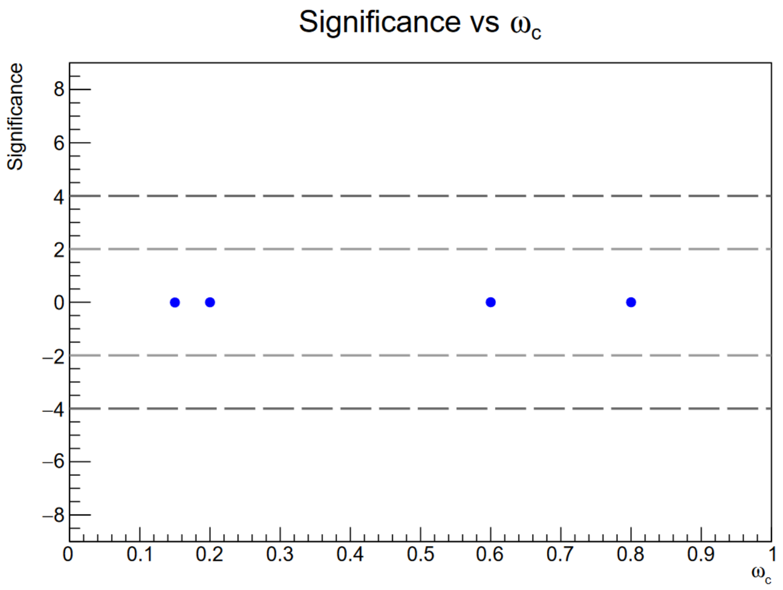


Figure 8.25: The Barlow Significance as a function of the normalisation used in the analysis. The normalisation was evaluated at $\omega_c = 0.388$ which is used as the optimal value to calculate the significance.

of the model for the mass radius of the bound proton. There are no obvious systematic effect due to the fact that the current measurements at CLAS12 do not have the statistical precision required to distinguish between models.

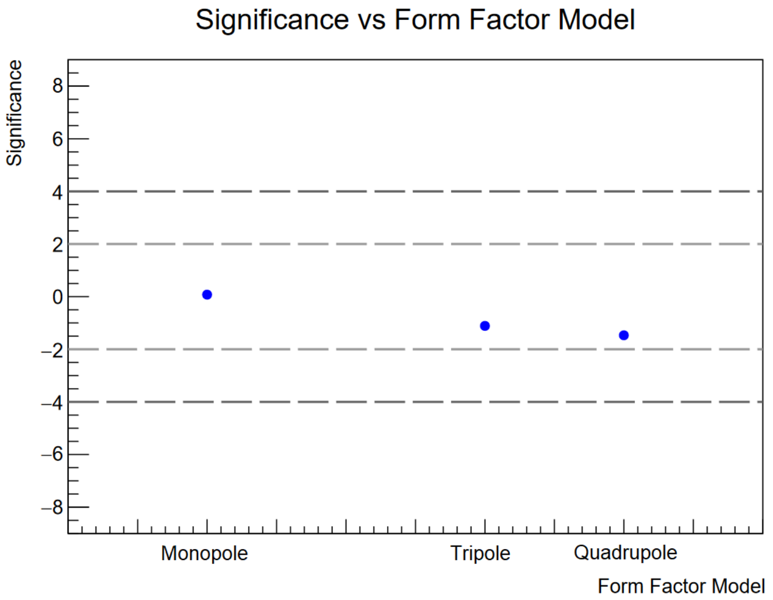


Figure 8.26: The Barlow Significance as a function of the form factor model, whether this is a monopole, tripole or quadrupole, with the optimal model taken as a dipole.

Chapter 9

Discussion and Outlook

Theoretical work since the start of the 21^{st} century has demonstrated the importance of J/ψ near-threshold photoproduction. A large component of the mass of the proton, an ongoing puzzle for modern nuclear physics, is thought to be due to the trace anomaly of the energy momentum tensor (EMT) of QCD [8] and the gluonic content of the nucleon [44]. Predictions of the total [44] and differential cross section [49] of J/ψ near-threshold photoproduction support the assumption of a dominant two-gluon exchange as the dominant production mechanism near the J/ψ production threshold. This assumption then allows one to relate the J/ψ differential cross section to the trace anomaly of QCD [16, 53, 55]. So called gluonic Gravitational Form Factors (GFFs) that describe the mechanical properties of the nucleon, such as the mass, pressure, and shear distributions of gluons within the nucleon [58], are defined from the matrix elements of the EMT. A scalar GFF describes the mass radius of the nucleon [57]. The GFFs can be estimated from the J/ψ near-threshold differential cross section [57, 58, 61]. Due to the extreme weakness of the nucleon gravitational field and the current experimental limitations for studying graviton-proton scattering [13, 57], J/ψ near-threshold photoproduction offers an unrivaled approach to studying the mechanical properties of the gluonic content of the nucleon.

The assumption that the mass radii and GFFs can be measured from J/ψ photoproduction data is heavily model dependent. First off, some discussions [71, 80, 81] surrounding the validity of Vector Meson Dominance (VMD) place doubt on whether this can be used to model the J/ψ total and differential cross sections. However, so far the predictions from VMD models have been accurate in estimating the scale and slope of the cross sections, and other theoretical models such as holographic QCD also point to the relationship between the trace anomaly of QCD and J/ψ near-threshold photoproduction [16]. Another critical assumption is that of a dominant two-gluon exchange. Some studies have shown that a three-gluon exchange mechanism may be dominant near-threshold, with the two-gluon exchange dominant further from threshold [27, 28], whilst some suggestion also exists that $J/\psi p$ photoproduction can be dominated by the production and decay of open charms states such as $\Lambda_c\overline{D}^{(*)}$ [70]. Future upgrades at JLab would enable a test of these assumptions [78] and to find photon energy ranges over which the two-gluon exchange dominance is accurate by measuring the charge parity naturality. The validity of the different theoretical assumptions remains at present hard to estimate. However the predicted importance of J/ψ photoproduction to modern particle and nuclear physics motivates the relevance of making these measurements at CLAS12.

The CLAS12 detector was used to take data with an electron beam interacting with a proton or deuteron target, producing J/ψ off of the free proton, the bound proton and the bound neutron. The analysis of J/ψ photoproduction at CLAS12 starts with the identification of final state particles such as the recoil nucleon on which J/ψ was produced and the electron positron or di-muon pair produced in the J/ψ decay. Requiring a missing mass squared close to zero allows to enforce the exclusivity of the event, with cuts on Q^2 allowing to select quasi-real photoproduction events. Along the way, new tools were developed towards the analysis of J/ψ photoproduction at CLAS12, such as a machine learning based muon identification algorithm which demonstrates the ability to detect and identify muons at CLAS12 despite not having a dedicated muon detector. Once the final state particles and J/ψ events were selected, simulation studies and a normalisation based on the known Bethe Heitler cross section allowed to calculate the J/ψ total and differential cross sections.

Some work remains to bring these measurements to publication. First, the present measurements are limited by low statistics. Improvements in the reconstruction algorithm at CLAS12 will soon enable a roughly 40% increase in statistics, with minimal changes required to the analysis. No systematic effects were identified in the analysis, as it is currently dominated by the statistical uncertainty, but the framework to identify and account for systematic effects has been developed. Another puzzling aspect of the analysis is the difference in the fit mean and width of the J/ψ peak when comparing J/ψ produced on the bound proton and on the bound neutron. A more rigorous study of the model assumptions used to fit the di-lepton invariant mass is required and might shed light on these discrepancies. Given that an increase in statistics will help draw more certainty from this study it was decided to wait for the aforementioned improvements to the CLAS12 reconstruction algorithm.

At present the main point of contention with the analyses presented in this thesis is the large normalisation factor. This factor is estimated by comparing the known Bethe Heitler cross section to CLAS12 data. It must be stressed that the normalisation factor used for the results shown in this thesis is still being investigated and will be updated soon. The uncertainty on the normalisation will also be established. The low statistics for Bethe Heitler at CLAS12 suggest a large associated uncertainty which is not being taken into account. As such, comparisons of the total cross section to previous measurements should be taken with some healthy scepticism. They demonstrate that a normalisation factor on the order of $w_c = 0.388$ does adequately correct the total cross section measurements at CLAS12 but a more in depth understanding of this normalisation is required to finalise these measurements. The good agreement of the mass radius measured at CLAS12 and at other experiments also suggests that there is no t dependency for the normalisation. The large normalisation suggests a sizeable error in the way that the detection and reconstruction efficiencies are currently simulated at CLAS12. Some work has already identified issues with the background merging that is used to simulate the reconstruction efficiency. Higher statistics will enable a more precise measurement of the normalisation factor, which may then help discern differences between different channels and the kinematic dependence of the inefficiencies. The added statistics coming from the improvements in the CLAS12 reconstruction will help towards better understanding the normalisation factor. In addition, future steps will involve studying two pion photoproduction that will provide a much more precise understanding of the normalisation due to greatly increased statistics. This will further improve the current understanding of the normalisation.

The J/ψ total cross section measured on the bound proton, bound neutron and free proton in both the decay to an electron positron pair and to a di-muon pair are shown in Figure 9.1. Although previous measurements at the JLab GlueX and Hall C collaborations have already measured the cross section on the free proton, the results from CLAS12 constitute a first measurement of the J/ψ total cross section on the bound proton and on the bound neutron in the near-threshold region. The good agreement between the total cross section on the bound proton and bound neutron supports isospin invariant production mechanisms such as a two- or three-gluon exchange. At present there are no predictions for open charm production of $J/\psi n$ and measuring the J/ψ photoproduction cross section on the neutron may allow for further tests of the J/ψ production mechanism. This thesis also presented a first measurement of the differential cross section as a function of t of J/ψ photoproduction on the bound proton and bound neutron. Good agreement between both channels suggests similar GFFs and mass radii for the bound proton and bound neutron as shown in Figure 9.2. Combining the information from both the bound proton and bound neutron allows to extract the mass radius of the bound nucleon with higher statistical precision than the individual measurements. With an increase in statistics this may allow for a higher precision comparison of the mass radius of the bound nucleon and free proton which may give an indication of EMC type effects.

Some suggestion in literature exists on the possibility of measuring the mass radius of the nucleon from lighter vector meson photoproduction [62, 63], such as ω or ϕ photoproduction. These results are also shown in Figure 9.2 in the energy bins closest to that available at CLAS12 with CLAS12 measurements of the mass radius of the bound and free proton from ϕ meson photoproduction. The measurement of the bound proton mass radius from ϕ photoproduction is also a first measurement. Figure 9.2 shows a notable difference between the mass radius

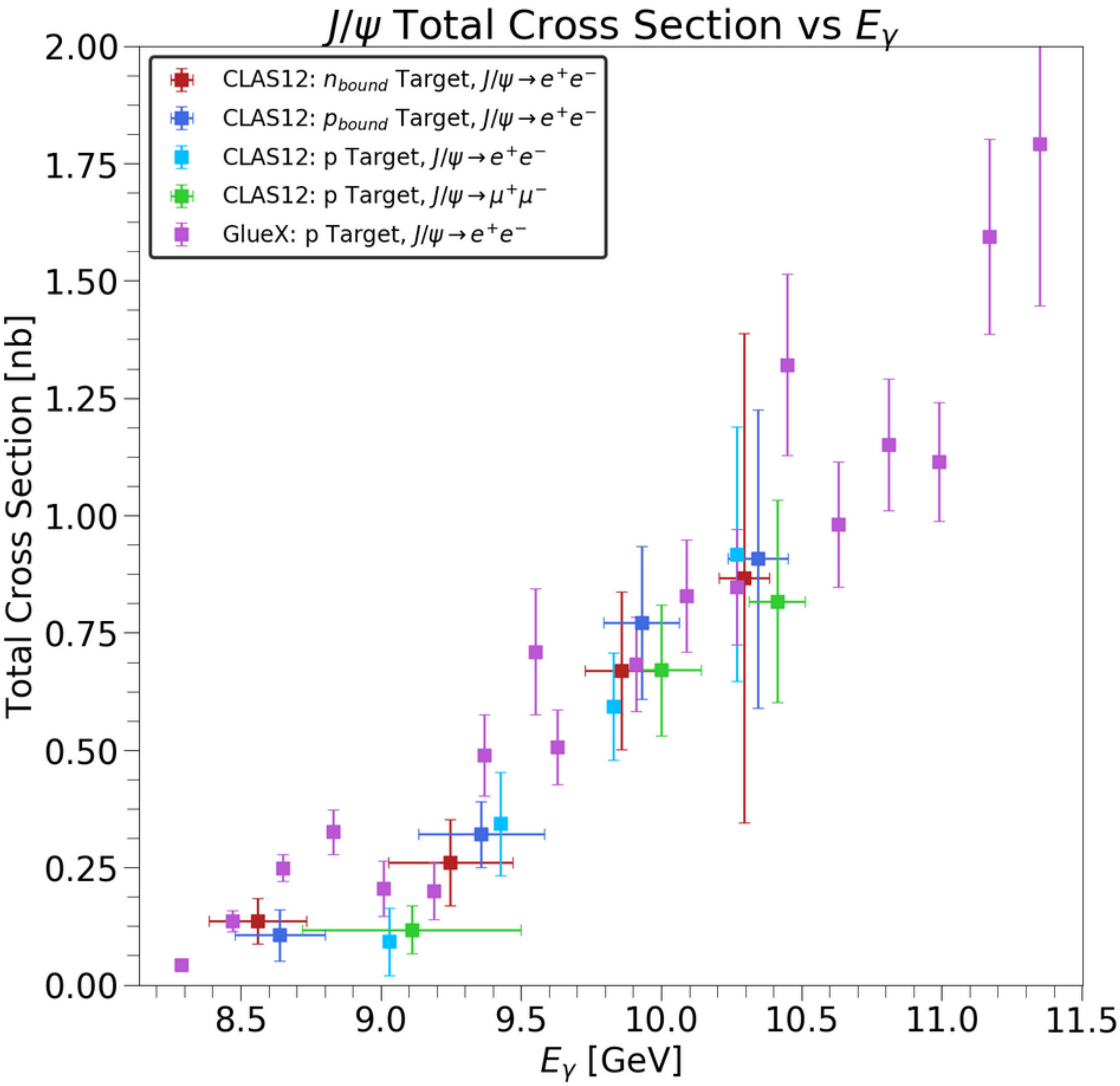


Figure 9.1: The J/ψ total cross section measured at GlueX (magenta) on the free proton in the J/ψ decay to an electron positron pair, compared to the CLAS12 measurements for J/ψ produced on the free proton decaying to e^+e^- (cyan) or $\mu^+\mu^-$ (green) and J/ψ produced on the bound proton (blue) and bound neutron (red) decaying to e^+e^-

measured from light vector meson photoproduction when compared to J/ψ photoproduction. One suggestion is that this is due to a dynamic response of the gluon contents of the nucleon when probed with colour dipoles of different masses [62,63]. Indeed, as discussed in Chapter 2, J/ψ near-threshold photoproduction requires the quark content of the nucleon to be in a compact fock state, and a similar effect could be observed for the gluonic content of the nucleon. It should be noted that the measurement of the mass radius from light vector meson photoproduction relies on the equivalence of a pomeron exchange with a two gluon exchange. The difference in mass radius when using J/ψ or light vector mesons as a probe might only be due to wrong assumptions on the production mechanism. More clarity on the matter would open an avenue

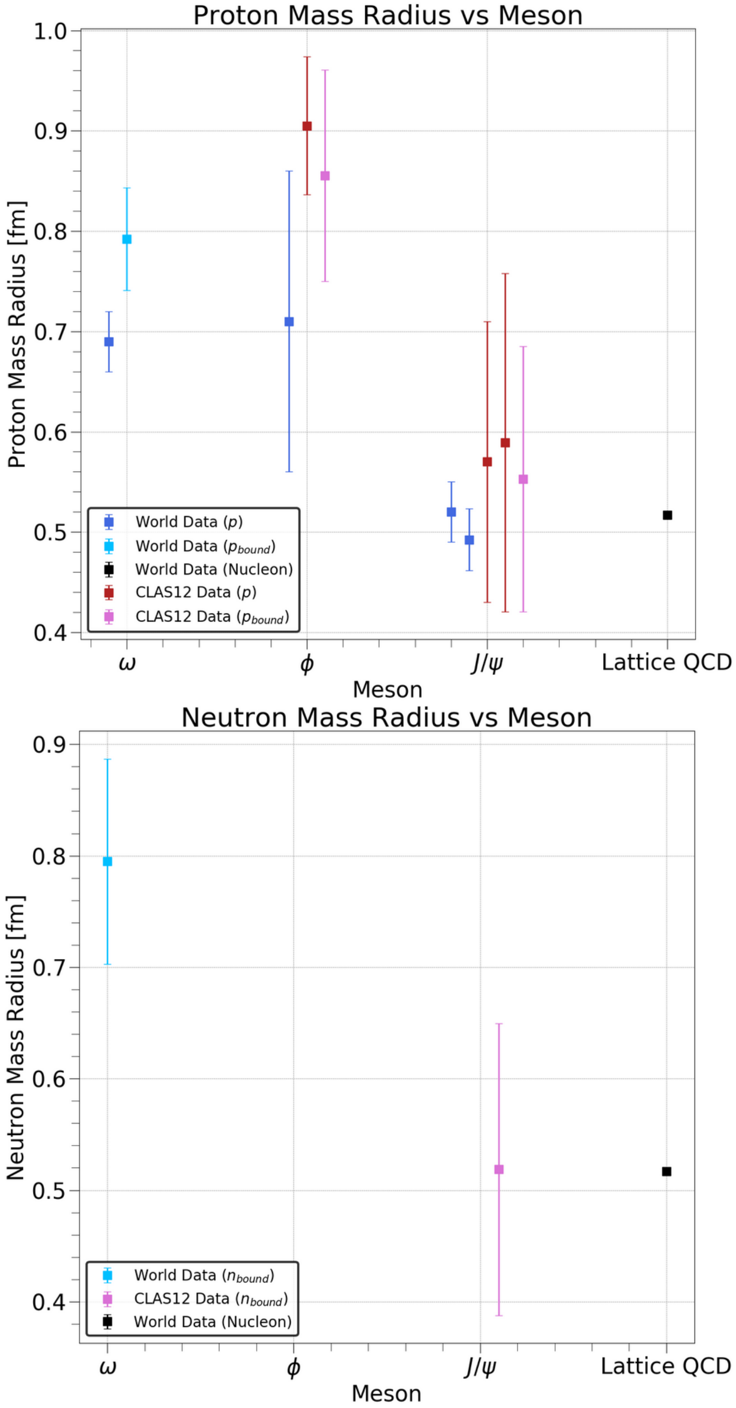


Figure 9.2: Top: The mass radius of the bound and free proton from the world data compared to the CLAS12 measurements for different meson probes and compared to a lattice QCD prediction for the nucleon. Bottom: The mass radius of the bound neutron from the world data compared to the CLAS12 measurements for different meson probes and compared to a lattice QCD prediction for the nucleon.

for higher statistics measurements of the mass radius of the nucleon. At present however, the results presented in this thesis constitute a first measurement of the mass radius of the bound proton and neutron from near-threshold J/ψ photoproduction.

Appendix A

Side Projects

This set of appendices aims to describe work that was done in parallel to the main objectives of this thesis. To avoid derailing the main narrative of the thesis these side projects are described here for the interest of the reader. The first appendix will describe how convolutional neural networks can be used to form an electron trigger at CLAS12. The second appendix gives a brief description of the *clas12root* and *chanser* toolkits. The third appendix will describe how machine learning methods can be used to create fast and accurate simulations. The final appendix will give an overview of the first preliminary steps that could be taken at CLAS12 towards a near-online J/ψ analysis.

A.1 A Deep Learning Level 3 Electron Trigger

The level 1 electron trigger currently operational at CLAS12 and described in Chapter 3 has an efficiency close 99.5% [109]. However, it was found to have a low purity. Essentially, although the trigger is very good at selecting events with an electron in one of the six sectors of the Forward Detector, it also introduces a lot of background due either to other particle types or to noise. The idea was then to use machine learning capabilities to improve the trigger by deploying a level 3 deep learning electron trigger that will filter out the noise introduced by the level 1 trigger. More detail on this can be found at [122].

The training dataset for the electron trigger was created by using the conventional offline reconstruction at CLAS12 to identify electrons in a given sector. The training data is then composed of events with an electron as the positive sample, and events without an electron as the negative sample, where instead there can be other particle types or simply noise. The AI trigger will use the coincidence of tracks in the drift chambers and hits in the calorimeters, as this will allow to detect charged tracks, with the calorimeter information allowing to identify electrons as described in Chapters 3 and 6. The drift chamber information is parsed into a 6 by 112 array representing the six superlayers in the DC that each have six layers or 112 sense wires.

The array is then filled with zero if a wire was not hit or 1/6 per hit per layer in a superlayer. The calorimeter information is parsed into a 6 by 72 array representing the three calorimeters, PCAL, ECIN, ECOUT and their three views, U/V/W. The array is then filled with the energy deposited in each strip, normalised by three as three GeV is the maximum energy that could be deposited in a strip [122]. Figure A.1 shows the two arrays used for the level 3 electron trigger.

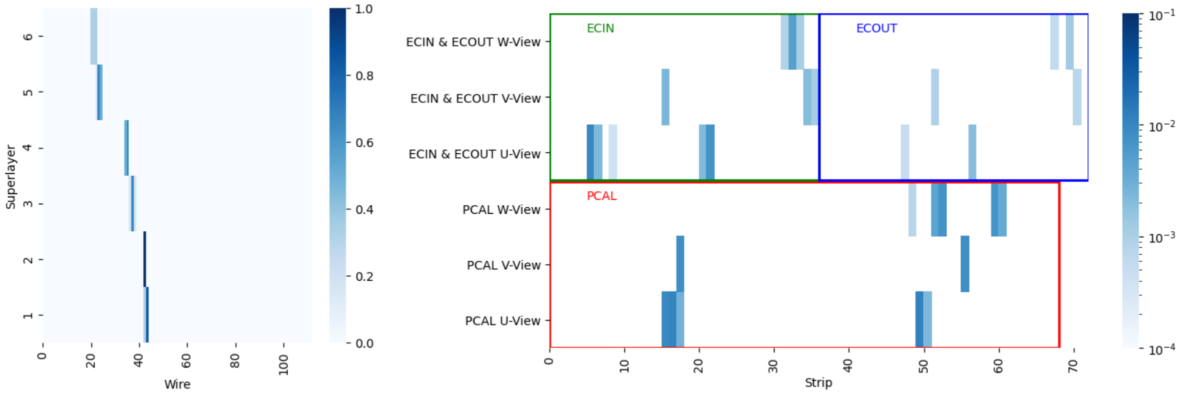


Figure A.1: Left: An example of a DC image containing an electron track in a given sector. The colour axis represents the number of layers with the same wire number, divided by six as normalisation. Right: an example of an ECAL image containing all ECAL hits in a given sector for a single event. The colour axis represents the energy deposited in each strip divided by three. Taken from [122].

To mimic data taken at high luminosity, the background merging process described in Chapter 4 was used to add noise to clean tracks identified using the CLAS12 reconstruction. Raw data taken at 50 nA for example can be simulated by adding background measured at 50nA. Data taken at 100 nA can then be simulated by adding the background twice. This allows to create raw data at a higher beam current than what was taken by CLAS12, allowing to study the performance of the level 3 trigger at higher luminosity [122]. To compare to how the trigger would perform without noise, a denoiser such as the one described in [106] is used to clean up the DC array. This then allows for a comparison of the performance of the trigger on noisy or clean data that could realistically be achieved online [122]. The process of noising and denoising the data is exemplified for the DC array in Figure A.2.

The choice was made to use a convolutional neural network (CNN), as described in Chapter 5, as these are the go to algorithm for image like data such as the arrays of Figure A.1. The information from the DC and ECAL is passed to separates sets of convolutional layers [122]. This then creates a modular architecture for the trigger, allowing to add other detectors if needed. For example, the trigger could be modified to trigger on charged hadrons by adding a time of flight detector. The output of the convolutional layers is resized into one dimensional arrays which are stacked before being passed to a neural network composed of dense layers. The

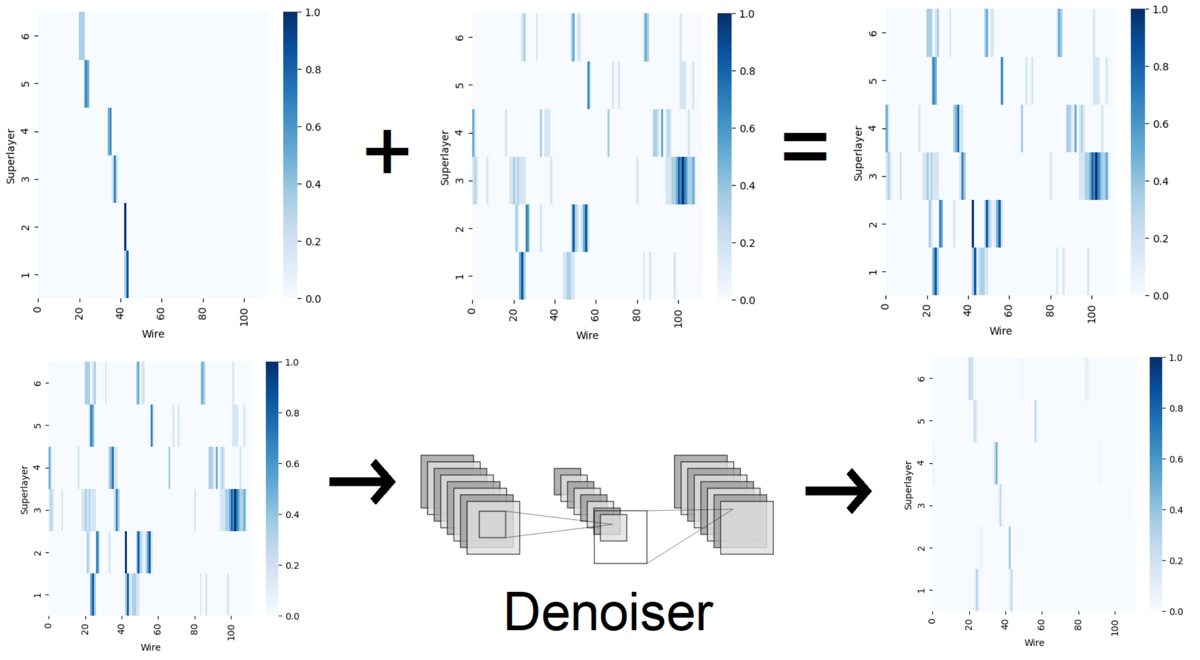


Figure A.2: The top row demonstrates the process of adding noise to a clean track to simulate raw hits as measured by the DC. The bottom row demonstrates how the de-noiser can be applied to the raw hits to return a cleaner track. Taken from [122].

neural network then outputs the trigger decision in the form of the response. Figure A.3 shows a schematic of the CNN architecture.

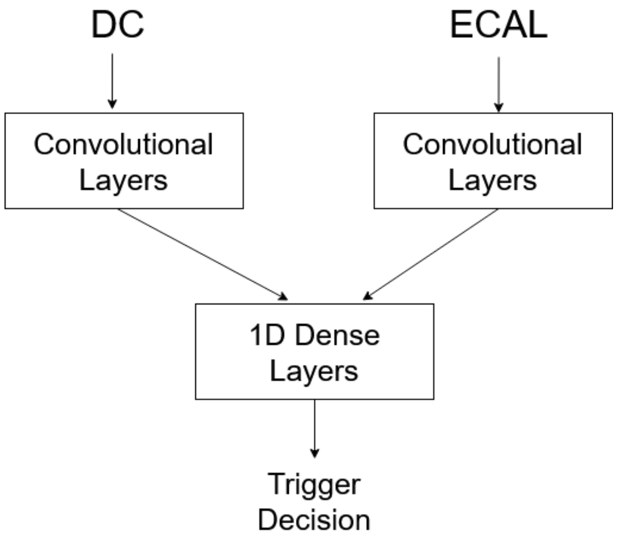


Figure A.3: The architecture chosen for the AI trigger CNN. Taken from [122].

The process of applying a threshold to the response and evaluating the efficiency and purity of the AI trigger as a function of this threshold on the response was already discussed in Chapter 5.

Figure A.4 shows the purity and efficiency of the level 3 trigger compared to the level 1 CLAS12 trigger. The efficiency of the level 1 trigger is assumed to be above 99.5% as it is not possible to estimate the efficiency offline given that events with an electron that were missed by the level 1 trigger would not be recorded. The efficiency of the level 3 trigger with respect to the level 1 trigger efficiency is stable above 99.5%. The purity of the level 1 trigger decreases as a function of beam current due to the fact that the algorithm it uses described in Chapter 3 is susceptible to noise. The level 3 trigger in comparison is robust to noise. This is exemplified first because adding the denoiser barely improves the performance of the trigger but also because it performs well as a function of beam current [122].

As the efficiency of the level 3 trigger is stable as a function of beam current, the level 3 trigger can then achieve a high data reduction rate at high luminosity. The data reduction is calculated as

$$DR = E_{AI}|(P_{AI} - P_{CLAS12})| \tag{A.1}$$

for E and P the efficiency and purity of the AI level 3 trigger and the CLAS12 level 1 trigger. Figure A.5 shows the data reduction achieved by the level 3 trigger. As shown, at high luminosity a sizeable proportion of the data can be removed, with most of this data being noise due to the high efficiency. Reducing the amount of background recorded is hugely beneficial in terms of saving on cost of storage and post processing times [122]. It also allows to decrease the online data rate which can then open the door to online analyses.

A final consideration is that the trigger must be able to keep up with the rate of data taking. Shown in Figure A.6, the level 3 trigger can reach up to a 100 kHz event rate when grouping the events into batches of variable size. CLAS12 has six sectors and so would require six predictions per event, which would then divide this prediction rate by six. However, the level 3 trigger will only be called on sectors where an electron was identified by the level 1 trigger, decreasing the number of predictions required per event. The rates shown in Figure A.6 were measured on a GPU with a Nvidia GeForce RTX 2080 Ti graphics card with 11 GB GDDR6 RAM and 4352 CUDA core. The suggested deployment of the level 3 trigger would see several GPUs working in parallel to achieve a suitable event rate [122].

A.2 *clas12root* and *chanser*

In parallel to the work presented in this thesis, some work was done on the analysis software packages *clas12root* [111] and *chanser* [112] that were used to analyse the CLAS12 data used to measure the J/ψ photoproduction cross sections. This type of work is typically referred to as service work for the collaboration.

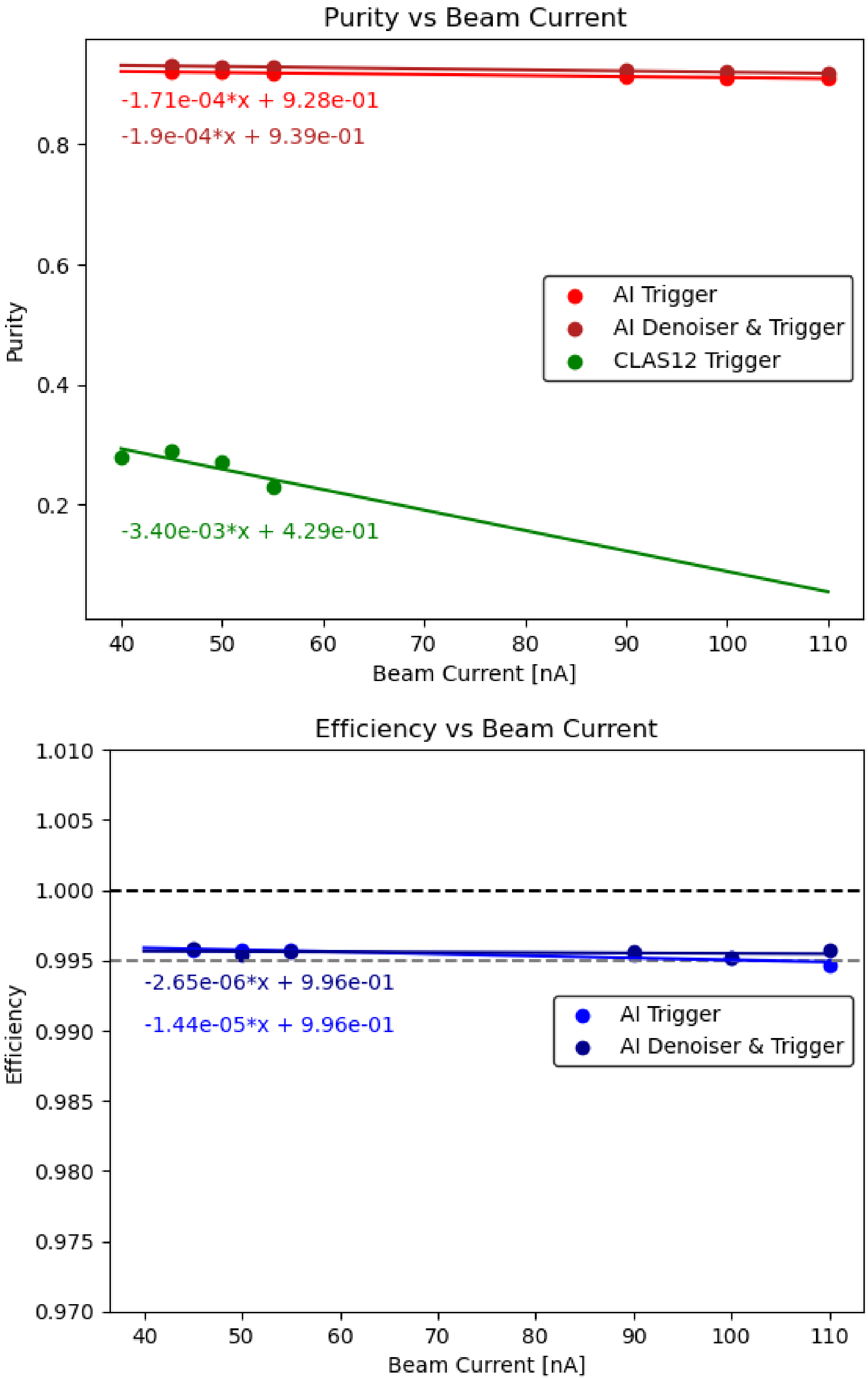


Figure A.4: Top: The purity of the AI (red) and traditional CLAS12 (green) triggers as a function of beam current. The AI trigger is compared with and without using a denoiser on the DC information. Bottom: The efficiency of the AI triggers as a function of beam currents. Taken from [122].

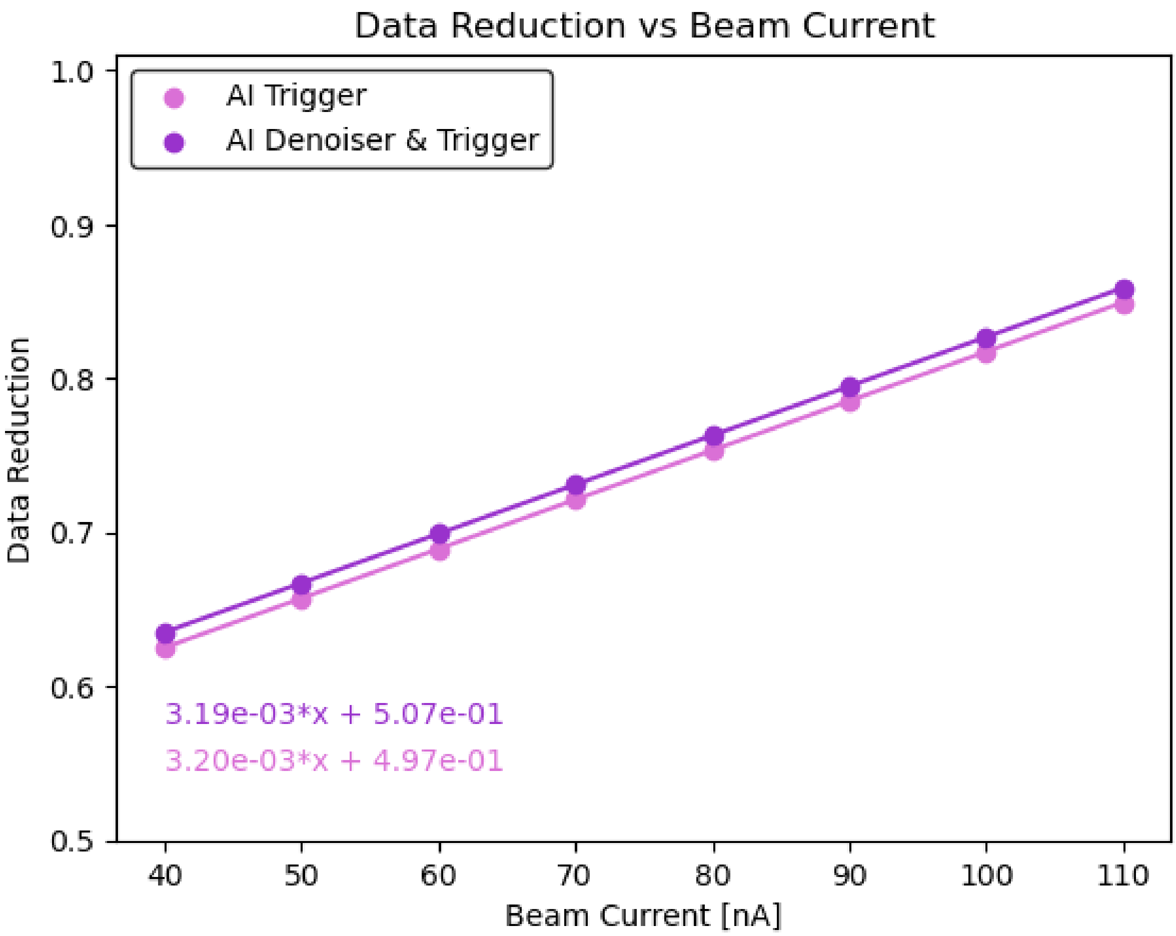


Figure A.5: The data reduction achieved by the level 3 trigger relative to the CLAS12 trigger. Taken from [122].

The goal of *clas12root* [111] is to facilitate access to the *HIPO* [110] file format. *HIPO* files are composed of banks typically corresponding to different detector subsystems but that can also be related to event wide information. Each bank then contains various entries, for example the event bank will contain the start time of the event. *clas12root* will then collate information from several subsystems into particles, allowing to easily access all the available information related to one particle in an event. *clas12root* also facilitates access to the particles in an event, for example enabling to access all electrons based on the event builder PID as described in Chapter 6. Finally, *clas12root* will loop over each event in a *HIPO* file, which is typically associated with one run during data taking. In such a way, *clas12root* allows to easily read all events in a run, and access the final state particles of each event for analysis.

A few features were added to *clas12root* as service work in parallel to this thesis. First, most analyses will only need a subset of all events in a run. A writer was added to *clas12root* to enable skimming of events during analysis. In the context of the J/ψ analysis, this could for example be done with loose requirements on the missing mass, Q^2 and the invariant mass, as

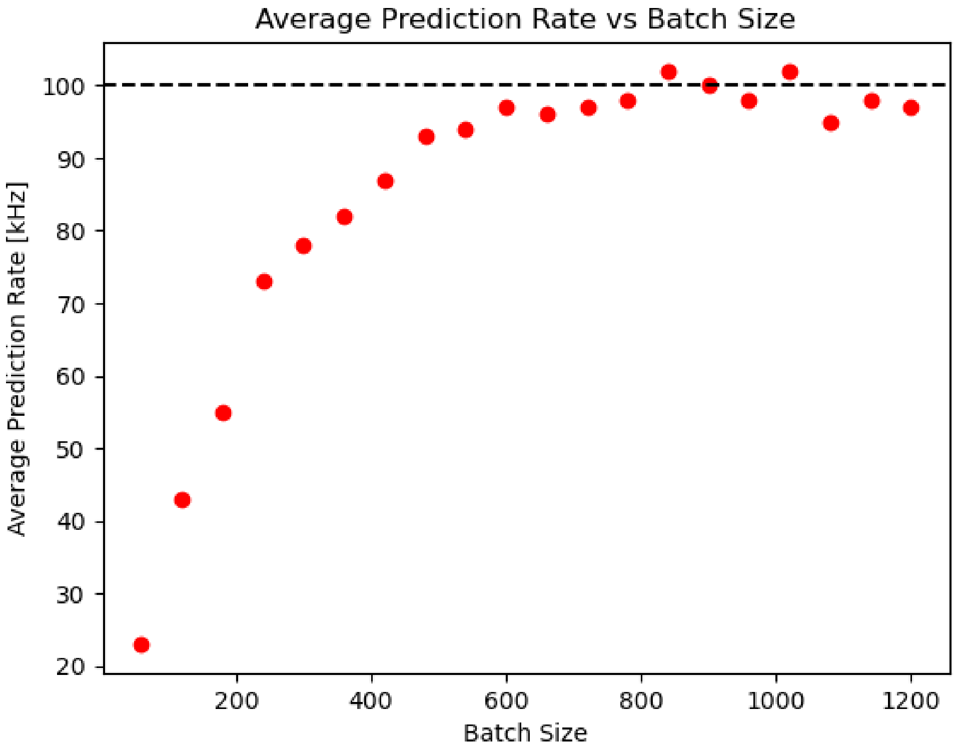


Figure A.6: The prediction rate of the AI trigger as a function of the batch size, for predictions grouped into batches. Quoted here are the average rates from 100 measurements. Taken from [122].

described in Chapter 7. This then allows to remove most of the data in a run as there a very few J/ψ events per run, allowing for much faster post-processing of each run. Second, access to the Run Condition DataBase (*RCDB*) [153] and the Quality Assurance DataBase (*QADB*) [154] was implemented in *clas12root*. The *RCDB* contains information about each run such as, for example, the type of target that was used or the beam energy and beam current. This information is often useful to access during analysis. The *QADB* contains information about the quality of each run, typically based on the ratio of the electron yield to faraday cup charge which should be constant. Implementing the *QADB* in *clas12root* allows to remove runs or events that were identified as low quality, and to count the charge that was accumulated in the runs that are kept for analysis.

The idea of *chanser* [112] is to build on *clas12root* and enable a modular analysis chain that can include various analysis procedures with little additional code and do most of the post processing for the analyser. The first step to building a *chanser* analysis is to define the required final state particles, for example electron positron and proton for J/ψ produced on the free or bound proton. The next step is to define the required topologies. For the J/ψ analyses presented in this thesis, the only topology required the two decay leptons and recoil nucleon, but an additional topology could be envisioned with the scattered electron detected in the Forward Tagger.

chanser deals with per event particle combitorials, taking each possible combination of final state particles as a separate event. These are then typically skimmed to only one combination per event with exclusivity cuts such as the ones described in Chapter 7.

chanser then allows to add data classes, cut classes and correction classes for each particle in the required topologies. The data class writes all required information to a ROOT file once *chanser* has finished processing all the *HIPO* files passed to it. The cut class applies some cuts to a given particle type, with the correction class applying corrections a given particle type. An example of a cut class would be the cuts on the energy deposition of muons detailed in Chapter 6. An example of a correction class would be the neutron path corrections detailed in Chapter 6. *chanser* also allows to add masking classes that typically remove unwanted particles. For example a masking class to remove duplicate neutrons as mentioned Chapter 6 was added to *chanser*, as was a class allowing to recover the momentum of electrons lost into radiated photons, and remove these secondary photons from the event. *chanser* also allows to train and deploy during the analysis machine learning classifiers from the *ROOT TMVA* software package, such as the muon or positron identification classifiers detailed in Chapter 6. Finally *chanser* will allow to record information relating to more than one particle, such as the missing mass of an event, in the same *ROOT* file as the information written by the data class.

Most of the analysis procedures detailed in Chapter 6 were coded into cut and correction classes. This then allows to quickly re-use the code and deploy it to another analysis. For example, the same code for J/ψ produced on the free proton in a hydrogen target and decaying to a di-muon pair can be used for J/ψ produced on the bound proton in a deuteron target with J/ψ also decaying to a di-muon pair. By adding the *RCDB* and *QADB* information, the code is automated to pick up the specific run conditions and ignore bad events such that the *chanser* analysis can be deployed on different run periods during different experiments.

A.3 Machine Learning Based Fast Simulation

Simulated data of high energy physics experiments is expensive to produce due to the high computing power required to simulate the interactions of particles with complicated detectors composed of many different subsystems. A typically expensive example is simulating electrons and photons in calorimeters. The showers that the electrons and photons produce in the calorimeters will require a lot of computation which then slows down the simulation. For some applications this can be problematic. For example, a large amount of simulated data is required when doing toy simulation studies and input/output tests, where a known simulated model is passed to the simulation, reconstruction and analysis software to ensure the known measurement is reproduced. This can then be very slow to produce with conventional simulation packages such as *GEMC* at CLAS12, as described in Chapter 4. Instead machine learning algorithms can speed up the simulation process to provide fast and accurate simulations. This is has been investigated in [155] and an overview will be given here.

The overall fast simulation scheme is shown in Figure A.7. The aim of the fast simulation is to recreate the acceptance and resolution effects modelled in a conventional simulation such as *GEMC*. The training process is separated into two stages: first deciding if a particle was successfully detected (acceptance); second, applying distortions to its reconstructed momentum components (resolution) [155]. The detection of each particle in the true event must be simulated independently, with separate models for each particle types. Training both the acceptance and resolution models relies on knowing the true information of the particles thrown at the conventional simulation. The true four-momentum of all particles thrown at the simulation and all particles accepted by the simulation is used to train the acceptance model. The true four-momentum of all particles accepted by the simulation and the four-momentum as reconstructed by the simulation is used to train the resolution model. During prediction, both models are chained to provide the full fast simulation [155].

The acceptance model relies on the fact that density ratio estimation can be used to treat the acceptance as a binary classification problem [155]. Given a dataset $\mathbf{x}$, split into two classes labelled $i = 0$ and $i = 1$, the model $\mathbf{c}$ is trained to successfully distinguish between both classes and outputs a response $C(\mathbf{x}) \approx p(i = 1|\mathbf{x})$ that the data sample falls into class $i = 1$. For the acceptance model, the labels correspond to the particles accepted by the simulation for label $i = 1$ and all particles thrown at the simulated for label $i = 0$. $\mathbf{x}$ contains the true four-momentum of the particles. The probability of accepting a particle $A(\mathbf{x})$ given its four-momentum components $\mathbf{x}$ is then [155]:

$$A(\mathbf{x}) = \frac{C(\mathbf{x})}{1 - C(\mathbf{x})} \tag{A.2}$$

The acceptance can therefore be simulated using any machine learning algorithm that can do binary classification, such as neural networks and boosted decision trees. In Figure A.8 the acceptance predicted by a neural network is compared to all thrown events (class $i = 0$ as described above) to those accepted by a conventional simulation (class $i = 1$). As shown, the neural network can accurately reproduced the acceptance of the conventional simulation. This can also be improved by adding another boosted decision tree that can reweight the classifier prediction to the conventional simulation [155], as shown in Figure A.9. This later steps allows for a more robust acceptance model that can allow for some discrepancy between the first neural network and the conventional simulation. In this example the first neural network was able to correctly reproduce the acceptance of the conventional simulation, but tests have shown that

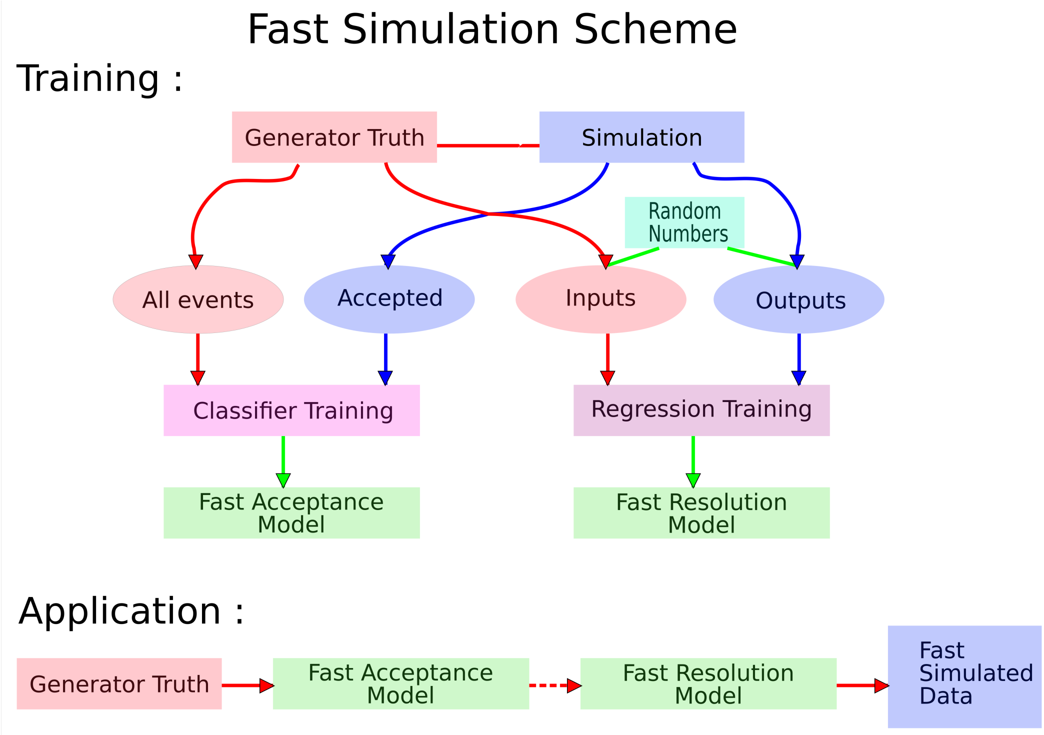


Figure A.7: Diagram of the fast simulation scheme. In training the true four momentum of the particle is compared to the simulated four-momentum. The fast simulation training is then separated into training the acceptance and the resolution effects. The acceptance uses all particles thrown at the simulations compared to the ones accepted in the simulation to create a model of the acceptance. The resolution uses the true four-momentum compared to the reconstructed four-momentum to create a model of the resolution effects. In prediction, the true four-momentum are passed to the acceptance model then the resolution model, the output of both is the fast simulated data. Taken from [155].

even for a poor initial prediction of the acceptance the subsequent reweighting step can allow for an overall satisfactory acceptance model [155].

The next step of the fast simulation scheme is to model the resolution effects of the conventional simulation. In this case a decision tree is used to predict the difference between the true and reconstructed four-momentum components given the true four-momentum components as inputs [155]. This is a regression task aiming to predict the error on the measurement of the four-momentum for each particle. Initially the decision tree will be overfitting to the data, meaning that it will predict exactly one value when passed the same values of momentum, θ and ϕ. Random inputs can be added to the four-momentum components to add some randomness to the training. This then allows the decision tree to predict more than one value when passed the same values of momentum, θ and ϕ. More randomness is introduced by training many decision trees with different random inputs and picking one decision tree at random during prediction, allowing to predict an almost smooth distribution of values given the same values of momentum, θ and ϕ [155]. The final outcome of the resolution simulation is shown in Figure A.10 showing

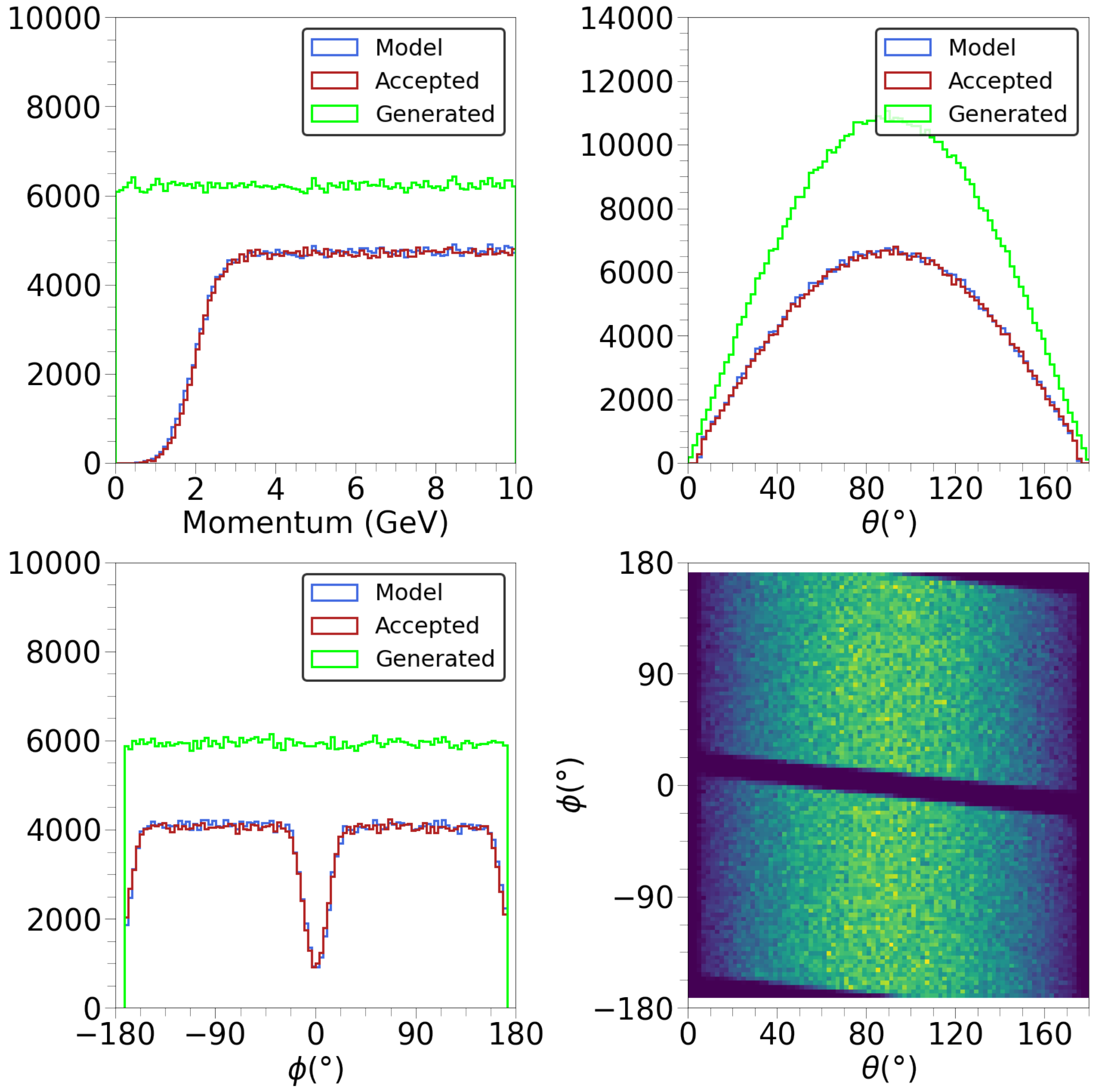


Figure A.8: Results of the acceptance simulation with a neural network shown for all four-momentum components and comparing all thrown events (Generated) to the ones accepted by the conventional simulation (Accepted) and the prediction of a neural network (Model). Taken from [155].

good agreement between the conventional and fast simulations.

The photoproduction reaction $\gamma + p \longrightarrow p' + \pi^+ + \pi^-$ was simulated using the conventional and fast simulations to exemplify how the fast simulation works for full reaction kinematics [155]. A separate model was trained for each individual particle. The differences between both simulations can be evaluated on the reconstructed distributions for the invariant mass of the three final state particles, W; the invariant mass of the two pions, $M(2\pi)$; the production angles in the centre-of-mass system ($\cos(\theta_{CM}), \phi_{CM}$); and the decay angles of the two pions in their

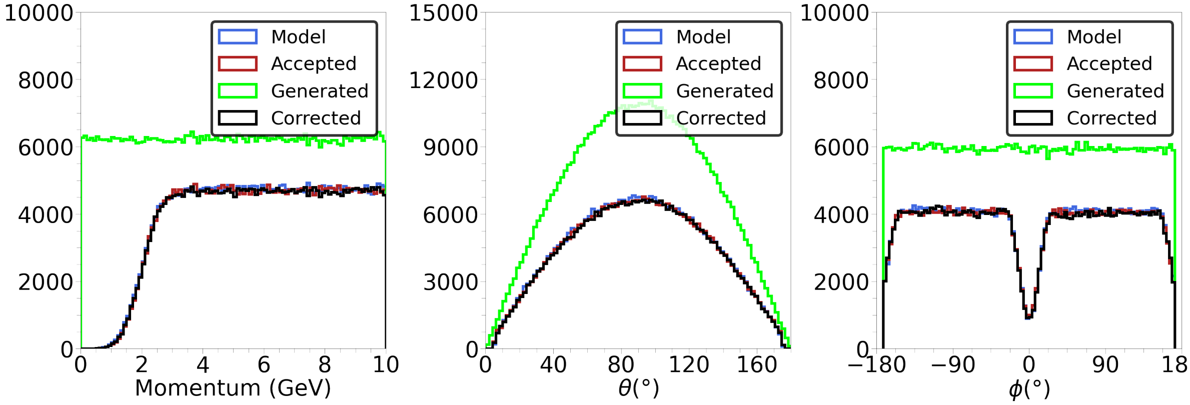


Figure A.9: Results of the acceptance simulation with an additional reweighting step shown for all four-momentum components (momentum, θ, ϕ) and comparing all thrown events (Generated) to the ones accepted by the conventional simulation (Accepted), the prediction of a neural network (Model), and the prediction of the neural network corrected by a boosted decision tree (Corrected). Taken from [155].

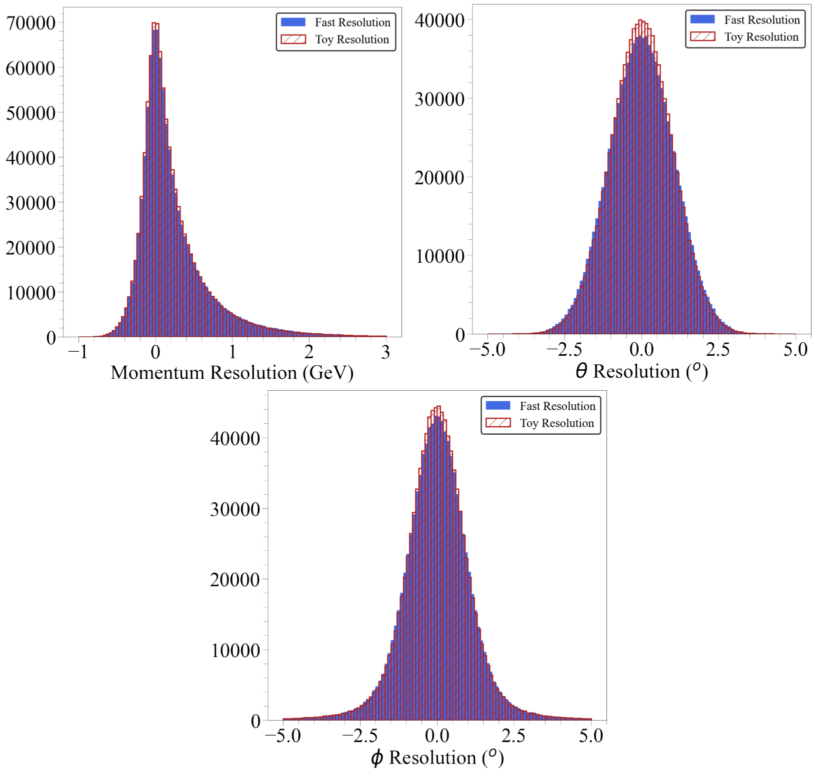


Figure A.10: The resolution of the fast simulation (blue) and conventional simulation (red) in momentum, θ, ϕ. Taken from [155].

combined rest frame (Gottfried-Jackson frame) $(\cos(\theta_{GJ}), \phi_{GJ})$. The distributions and ratio of the two simulations in shown in Figure A.11. The ratios for each distribution are all reasonably flat and close to 1. Slight deviations in the resolution distributions lead to slight deviations for very low counts in the variable distributions. The weighted average ratio in this test was found to be just over 0.99, giving an overall normalisation good to the 1% level. Most significantly, the fast simulation was able to run at 10 kHz [155] on a standard CPU which is several orders of magnitude faster than what could be achieved for a conventional simulation such as *GEMC*.

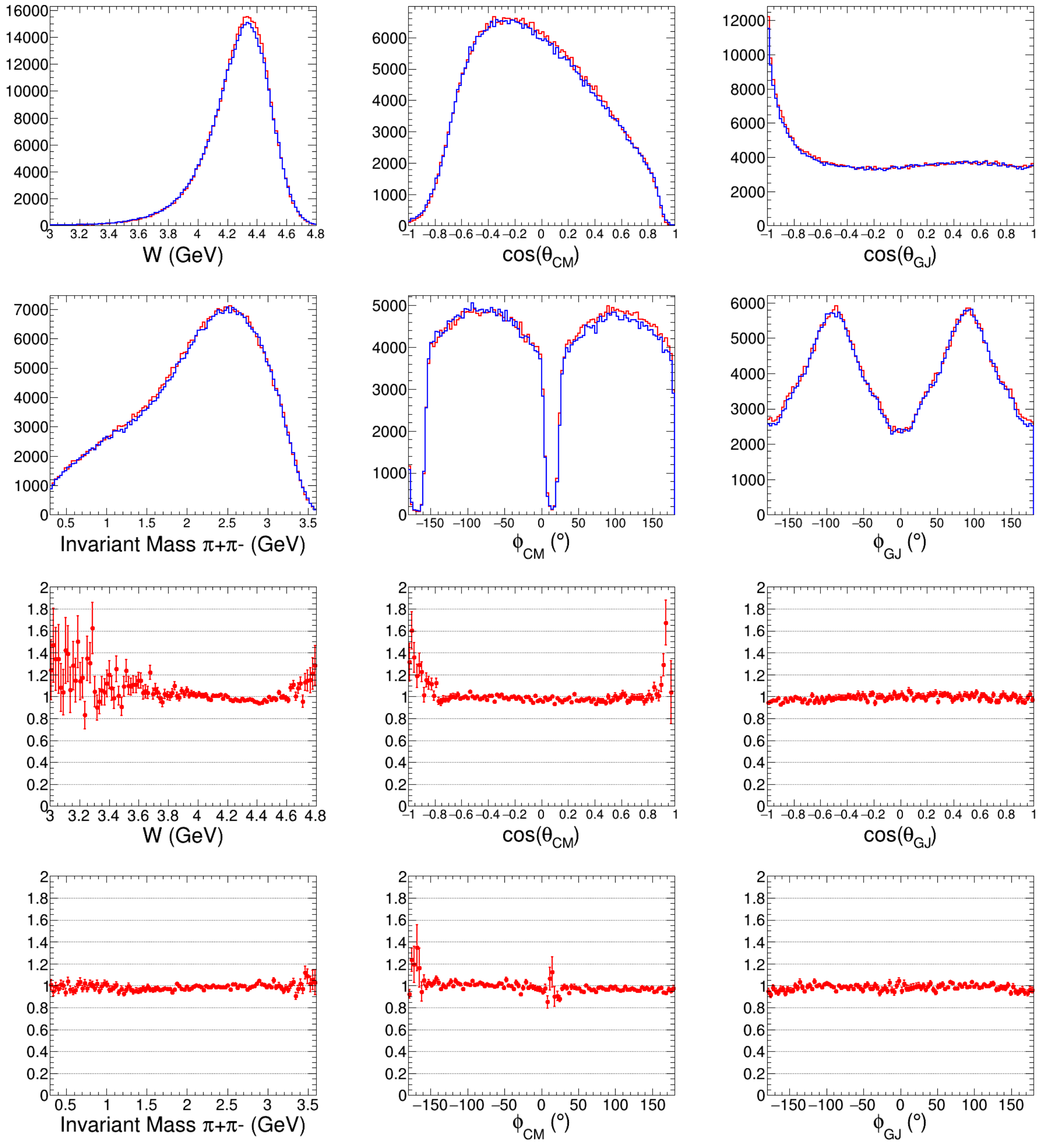


Figure A.11: Top: Accepted and reconstructed physics variables for the fast (blue) and conventional (red) simulations. Bottom: Ratio of conventional to fast simulations. The distributions show: the invariant mass of the three final state particles, W; the invariant mass of the two pions, $M(2\pi)$; the production angles in the centre-of-mass system ($\cos(\theta_{CM}), \phi_{CM}$); and the decay angles of the two pions. Taken from [155].

The studies in [155] were made using a toy simulation in lieu of the conventional simulation. Tests have been made seeing how the fast simulation can recreate the *GEMC* simulation at

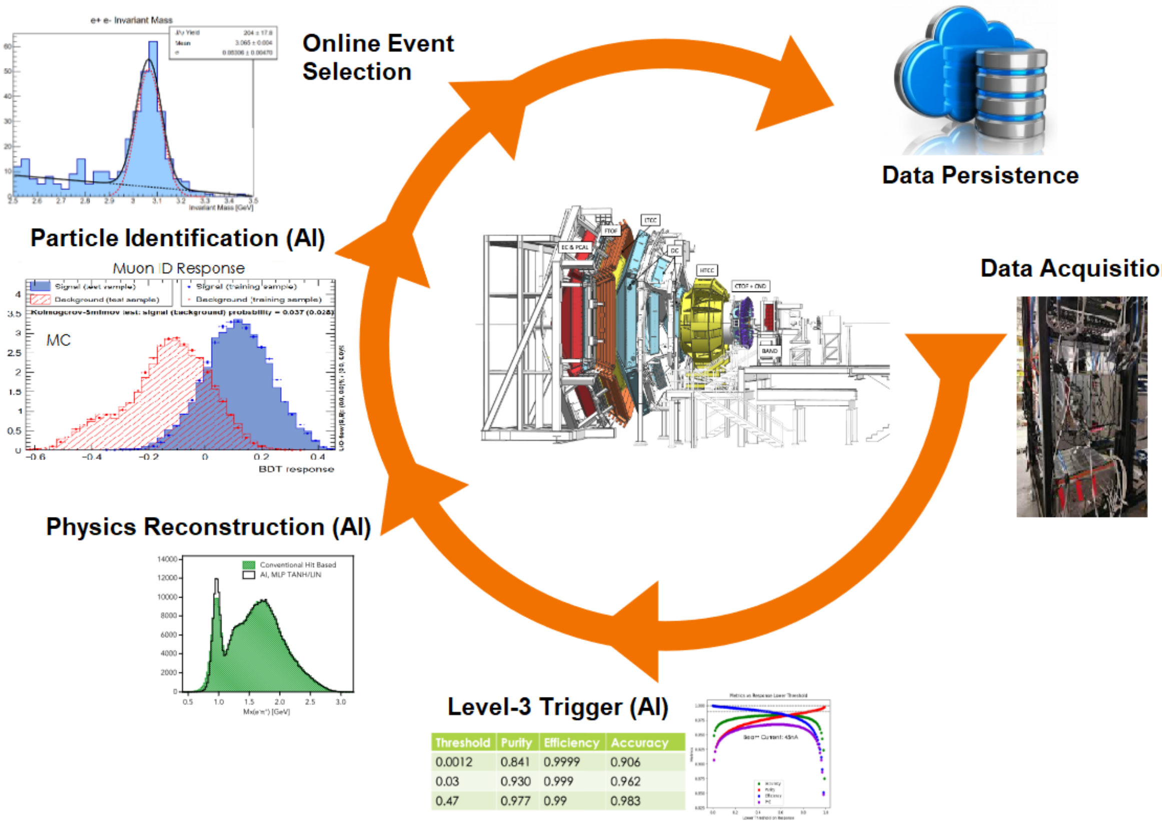
CLAS12. The main issue here is that the toy simulation assumed that for event made of N particles, the acceptance of the event factorises into the acceptance of the individual particles. In practice however there are some multi-particle correlations, for example due to overlapping tracks. These can be simulated correctly using an additional boosted decision tree such as the one used to reweight the acceptance. Some preliminary tests were also made investigating how the acceptance could be simulated directly from CLAS12 data as mentioned in Chapter 8. The main problem here is that although exclusive channels can be used to map the acceptance of individual particles, the multi-particle correlations are hard to estimate. On top of this, resolution effects mitigate the ability to map the acceptance as a function of momentum, θ, ϕ due to the non-negligible resolution introduced in measuring these quantities.

A.4 Towards a Near Online J/ψ Analysis at CLAS12

Near online analyses are useful due to the fact that they decrease the amount of time required to publish physics measurements and they enable online monitoring during data taking. Several of the tools described in this thesis can be combined to create a near online J/ψ analysis that would identify candidate J/ψ photoproduction events during online data taking. Although this has not yet been tested or deployed, the general idea of how this could be achieved at CLAS12 is described here.

Figure A.12 shows the steps that can be taken towards a near-online analysis at CLAS12. First the conventional CLAS12 DAQ streams the data towards a level 3 trigger as described in Appendix A.1. The conventional level 1 trigger is kept and events selected by the trigger are streamed to the level 3 trigger deployed on several GPUs in parallel. The level 3 trigger removes a lot of the background introduced by the level 1 trigger, which allows to save on costs of storage and post processing time, but also enables a lower data rate which opens the doors to the next steps of the online analysis.

The next step of the online analysis chain is reconstructing information from detector subsystems into physics variables, such as the four-momentum of final state particles. Machine learning based drift chamber reconstruction has already been investigated at CLAS12 [145], demonstrating that the momentum and angles of particles can be predicted from clusters in the drift chamber. An MLP was shown to be able to achieve the same resolution as conventional hit based tracking with a 500 times speedup [145]. The resolution of the MLP was worse than conventional time based tracking. However, this already demonstrates that the four-momentum of particles can be reconstructed online. From there, the conventional track matching can associate DC tracks to clusters in the ECAL, FTOF and HTCC. These latter subsystems may be reconstructed online using similar algorithms as to what was used for the level 1 trigger.



Threshold	Purity	Efficiency	Accuracy
0.0012	0.841	0.9999	0.906
0.03	0.930	0.999	0.962
0.47	0.977	0.99	0.983

Figure A.12: The steps towards a near-online J/ψ analysis at CLAS12. First the CLAS12 DAQ streams the data towards a level 3 trigger. This allows to reduce the data rate by removing a lot of background, allowing for the next steps of the analysis to be done online. Next, online reconstruction allows to reconstruct four-momentum of final state particles towards analysis. Particle Identification routines identify the final state particles, which are then grouped and selected as events. All this information is then written out for offline post processing.

The two next steps of the online analysis chain are particle identification (PID) and event selection. Chapter 6 described how either the conventional or machine learning based particle identification works. These algorithms are fast and accurate, but the conventional PID may require some calibration of detector subsystems. Machine learning algorithms for PID have been shown to improve on their conventional counterparts, and may be more robust to un-calibrated data. Chapter 7 has described how candidate J/ψ photoproduction events can be selected with cuts on the missing mass and Q^2, or using neural networks. Due to the potential bias that the neural network reaction identification tools can introduce, the conventional cut based event selection would probably be preferred.

All of this information is then saved along with the usual information from each detector subsystem. As mentioned in Chapter 7, online flagging of certain reactions such as J/ψ photoproduction could then allow for shorter timescales in producing physics measurements as only specific events could be targeted for reconstruction, removing an important timing overhead due to reconstructing background events. Furthermore, as detailed in Chapter 4, some calibrations depend on having robust PID and four-momentum reconstruction. Having this information already available as soon as the data is taken would allow for a faster calibration process. This also opens the door to performing calibrations during data taking. Although J/ψ is not the most useful reaction for online monitoring purposes as there are so few J/ψ events at CLAS12, other channels could be used to ensure that the expected rates for given reactions are reached by CLAS12 during data taking.

Some additional analysis steps will be required offline, explaining why this would only be a near-online J/ψ analysis and not fully online. These additional steps are, for example, acceptance corrections and normalisation that depend on simulated data. Appendix A.3 however has discussed how acceptance corrections could be estimated from data. Although this still requires additional work, it could one day allow to do acceptance corrections online. All the steps described in Figure A.12 are however realistically achievable. This will require further investigation, but opens up the exciting prospect of doing near-online analyses at CLAS12.

Appendix B

Invariant Mass Plots

In this appendix are shown the invariant mass plots for each bin of each channel of the total and differential J/ψ and ϕ photoproduction cross sections measured in this thesis. The J/ψ or ϕ peaks are fitted with a gaussian with parameters mean and σ. The J/ψ or ϕ yield gives the number of J/ψ or ϕ events fitted with the gaussian. The backgrounds are either fitted with a linear slope, with parameters slope and constant as in slope$\times$x+constant, or with a third order polynomial with parameters 1st, 2nd, 3rd order coef and an offset as in 1st$\times$x+2nd$\times$x^2+3rd$\times$x^3+offset.

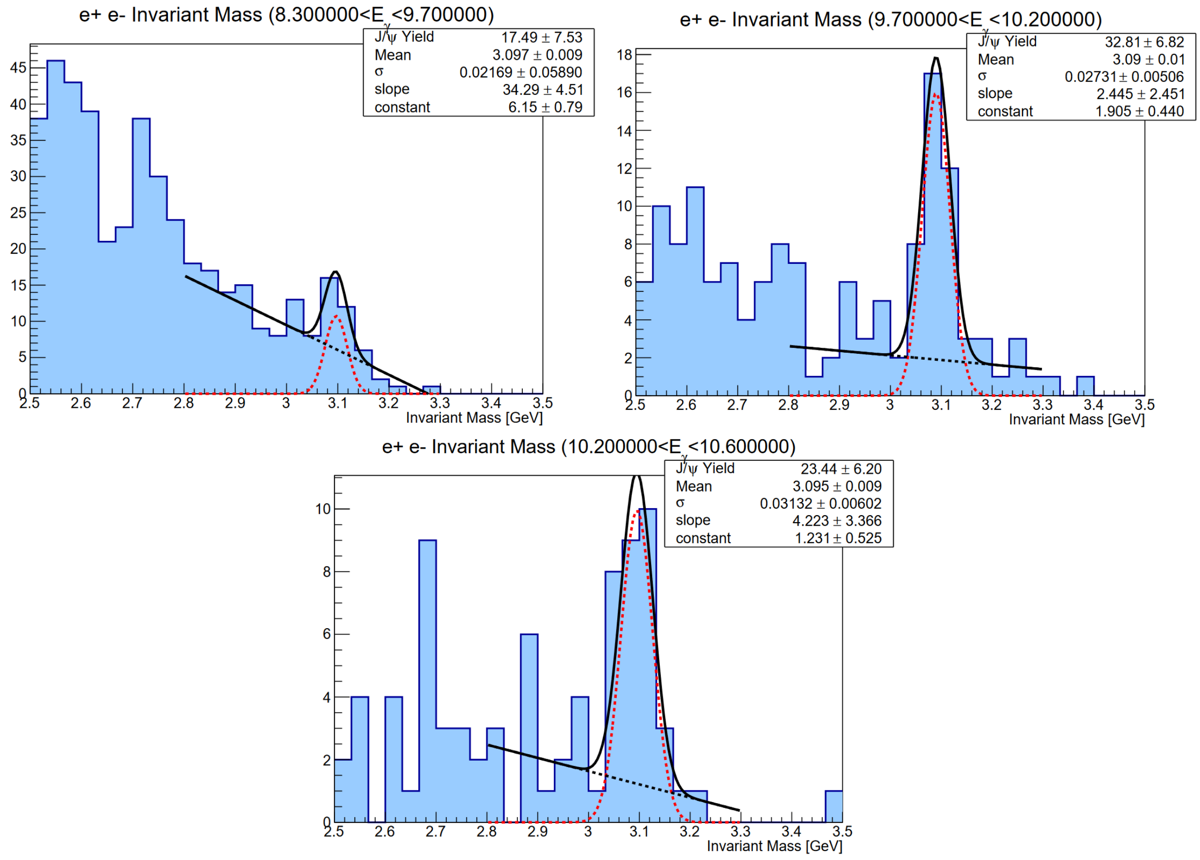


Figure B.1: The di-lepton invariant mass in the $ep \to (e')\mu^+\mu^- p$ channel when measuring the J/ψ total cross section produced on the free proton.

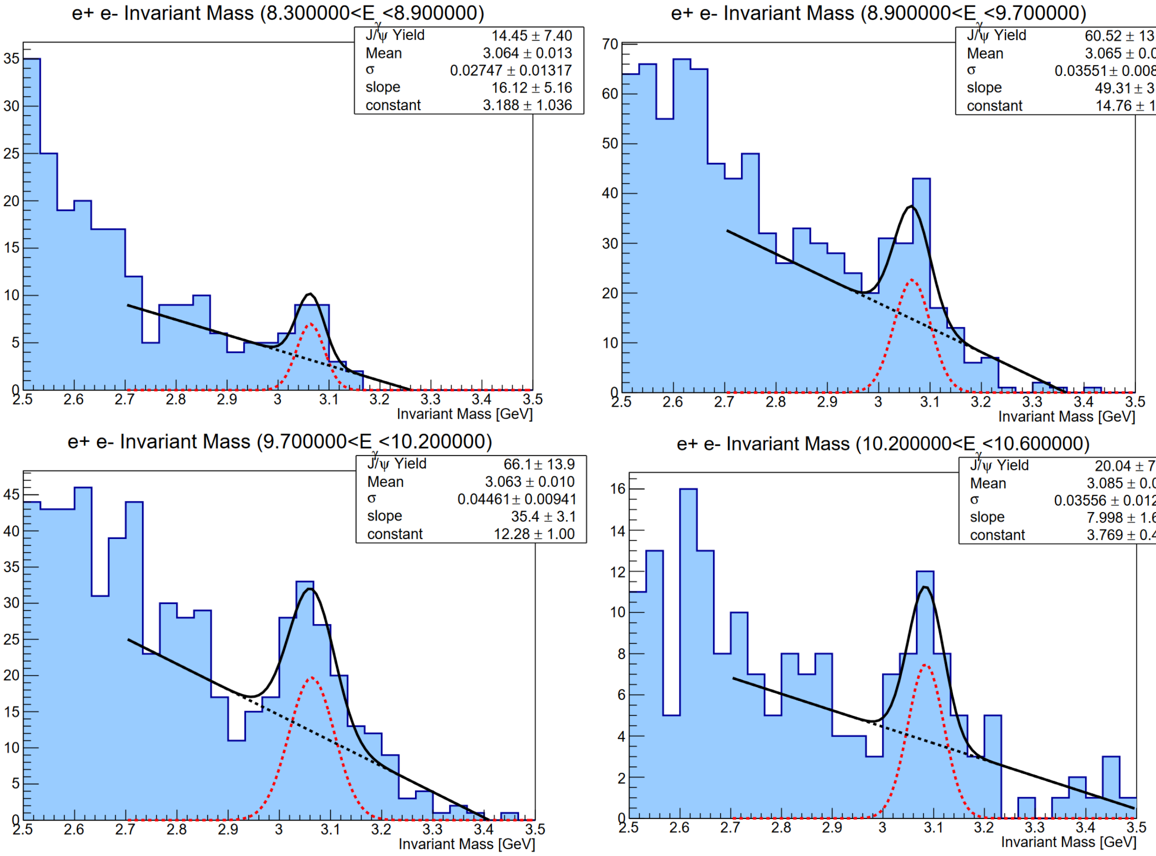


Figure B.2: The di-lepton invariant mass in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel when measuring the J/ψ total cross section produced on the bound proton.

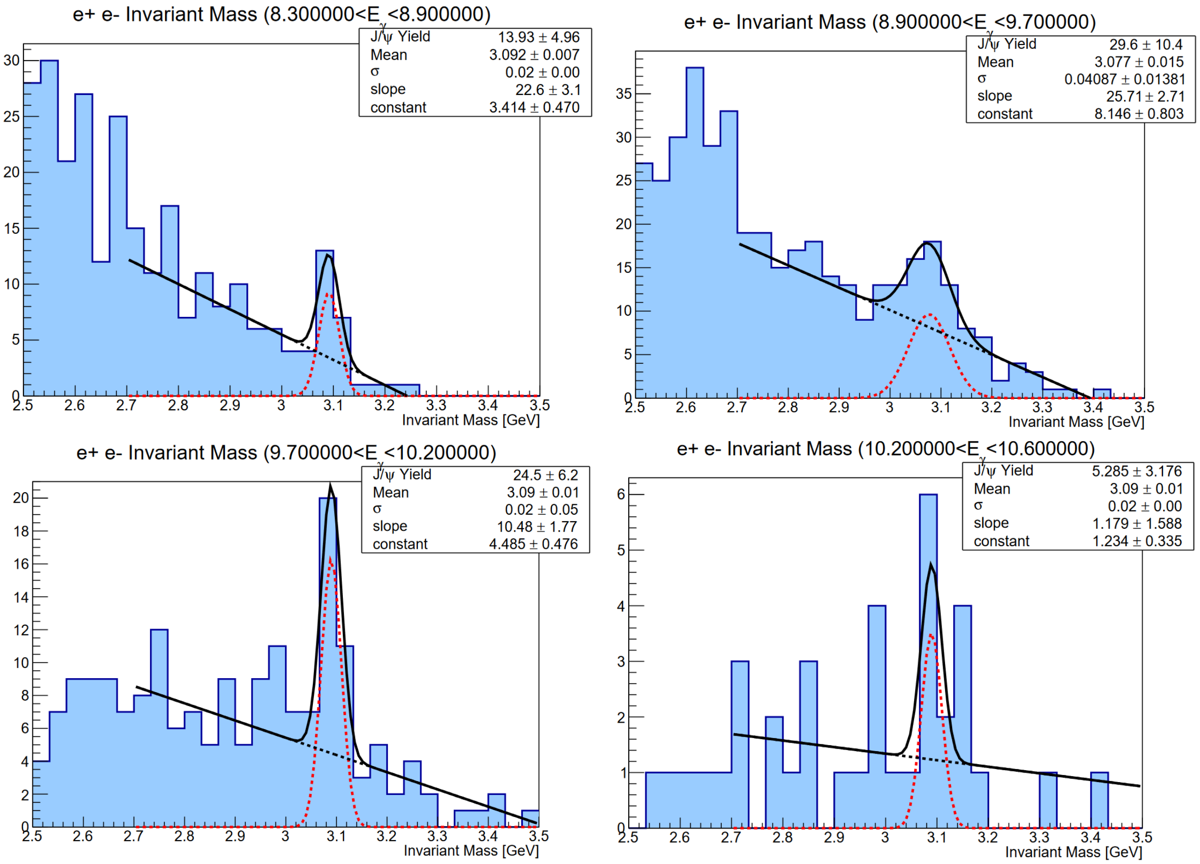


Figure B.3: The di-lepton invariant mass in the $en_{bound} \rightarrow (e')e^+e^-n$ channel when measuring the J/ψ total cross section produced on the bound neutron.

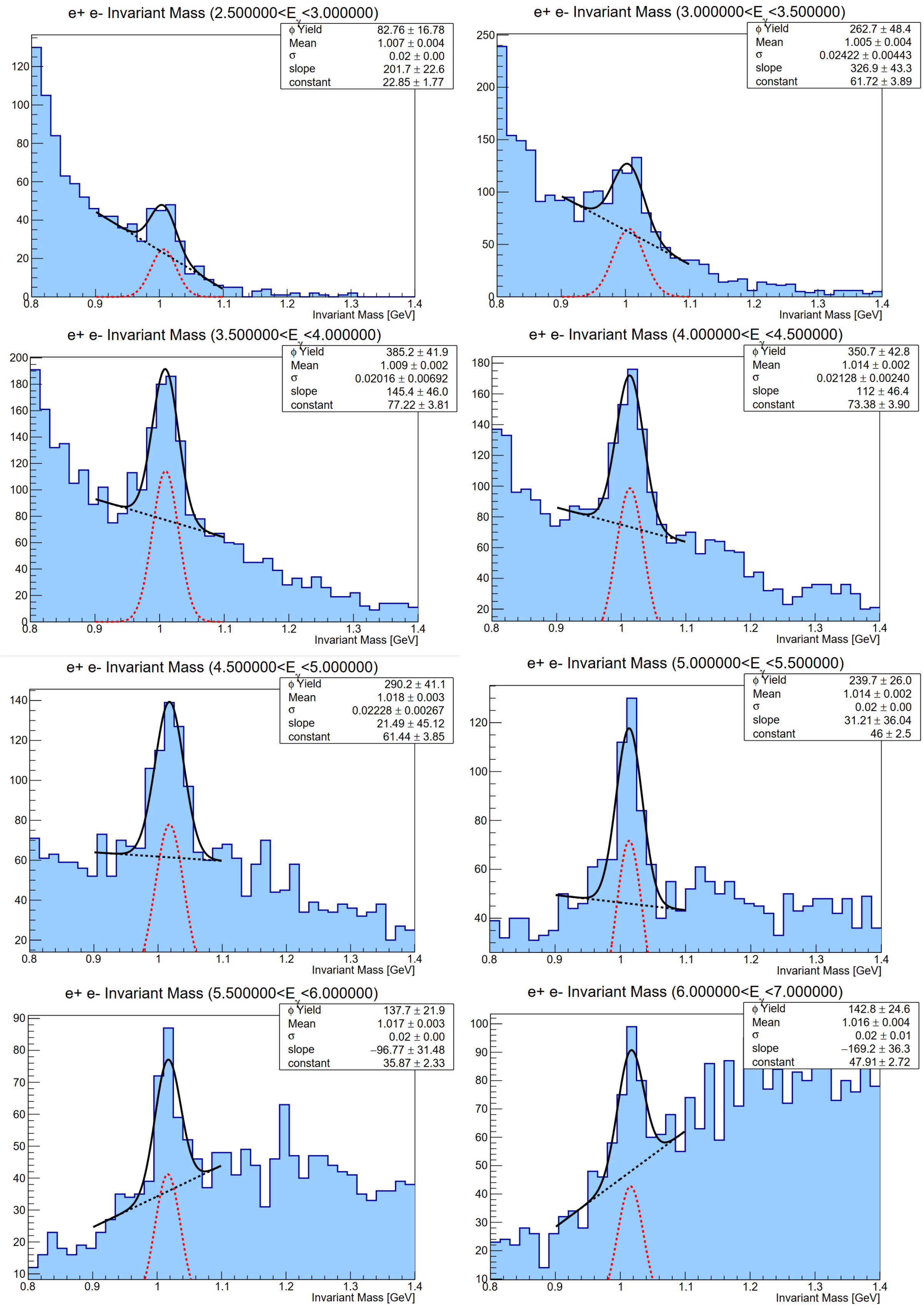


Figure B.4: The di-lepton invariant mass in the $ep \to (e')e^+e^-p$ channel when measuring the ϕ total cross section produced on the free proton.

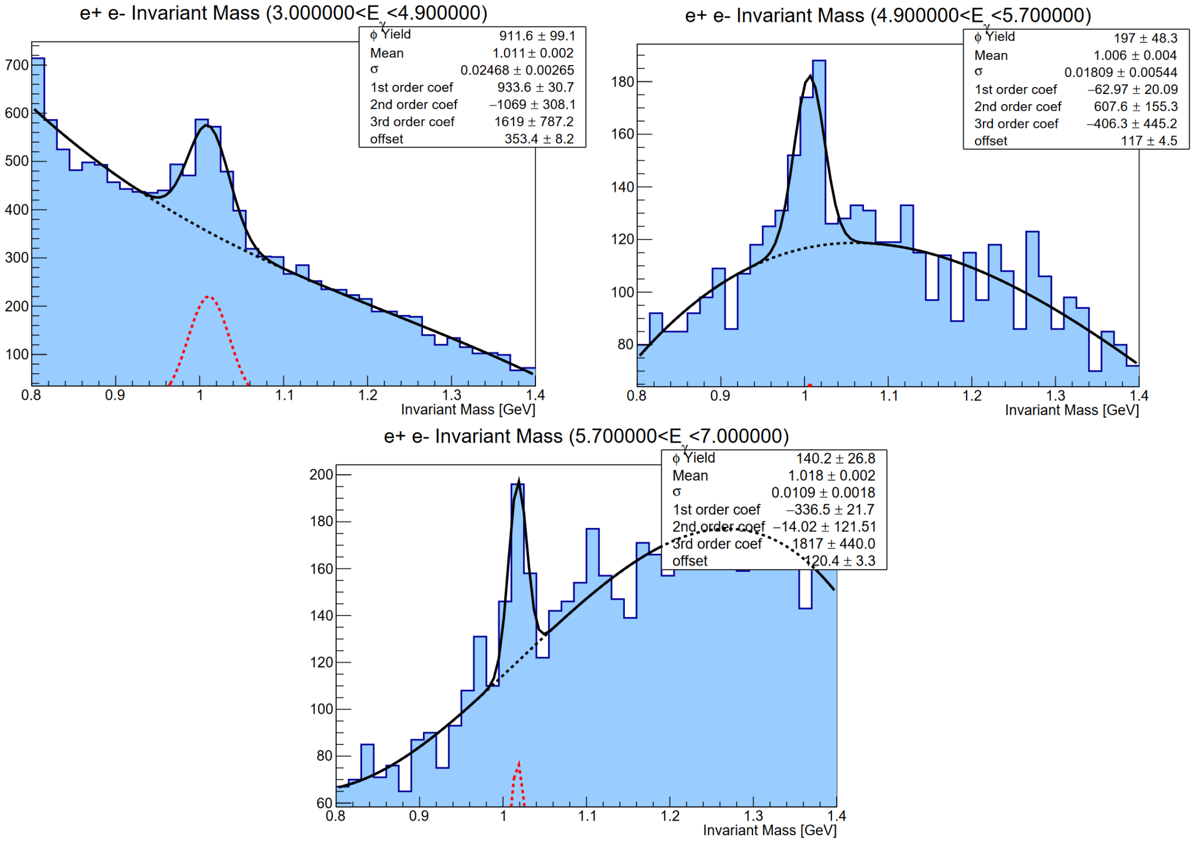


Figure B.5: The di-lepton invariant mass in the $ep_{bound} \to (e')e^+e^-p$ channel when measuring the ϕ total cross section produced on the bound proton.

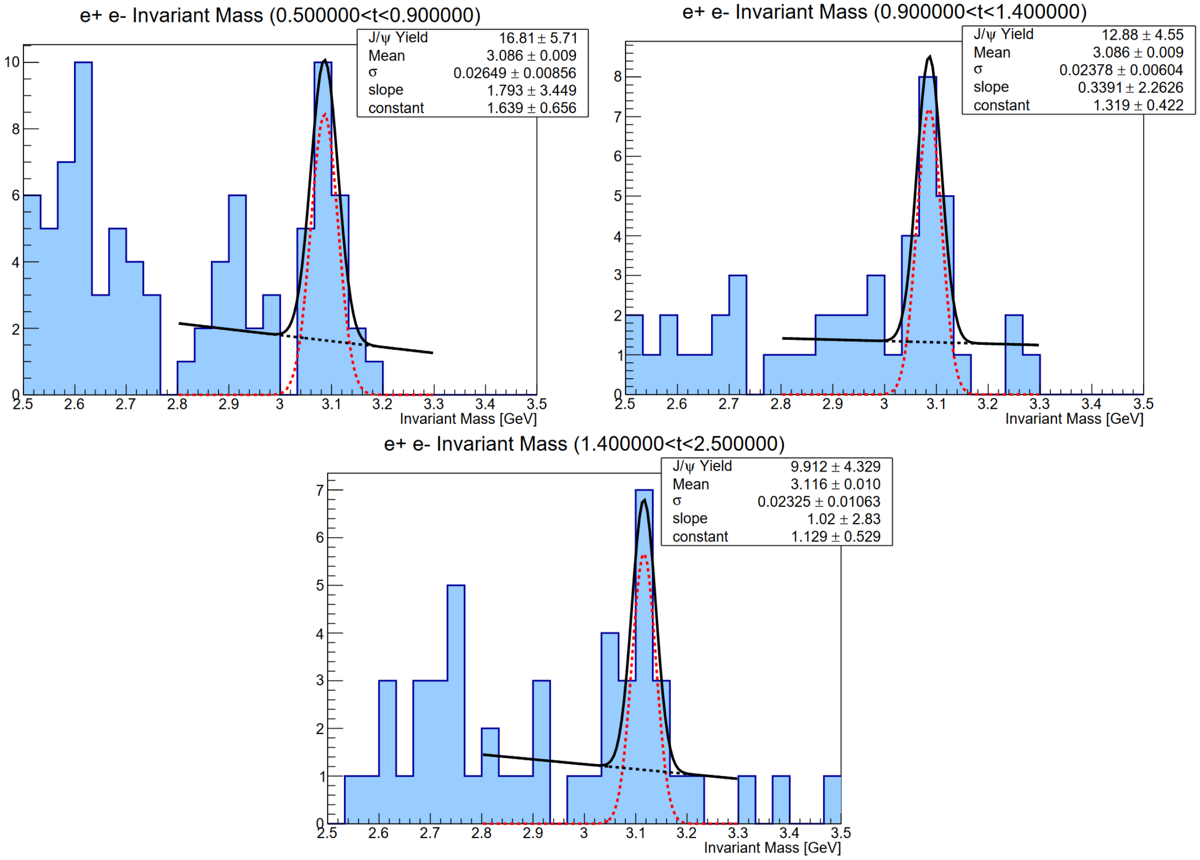


Figure B.6: The di-lepton invariant mass in the $ep \rightarrow (e')\mu^+\mu^- p$ channel when measuring the J/ψ differential cross section produced on the free proton.

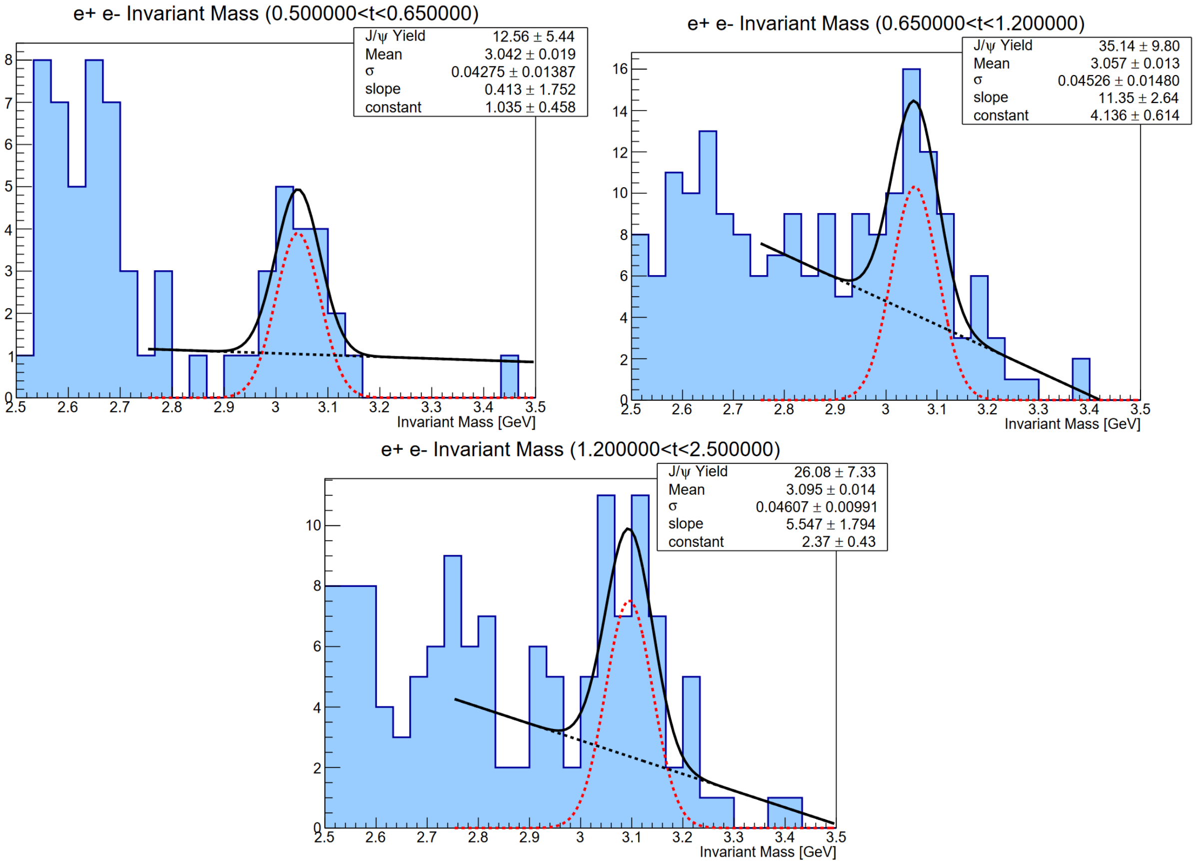


Figure B.7: The di-lepton invariant mass in the $ep_{bound} \rightarrow (e')e^+e^-p$ channel when measuring the J/ψ differential cross section produced on the bound proton.

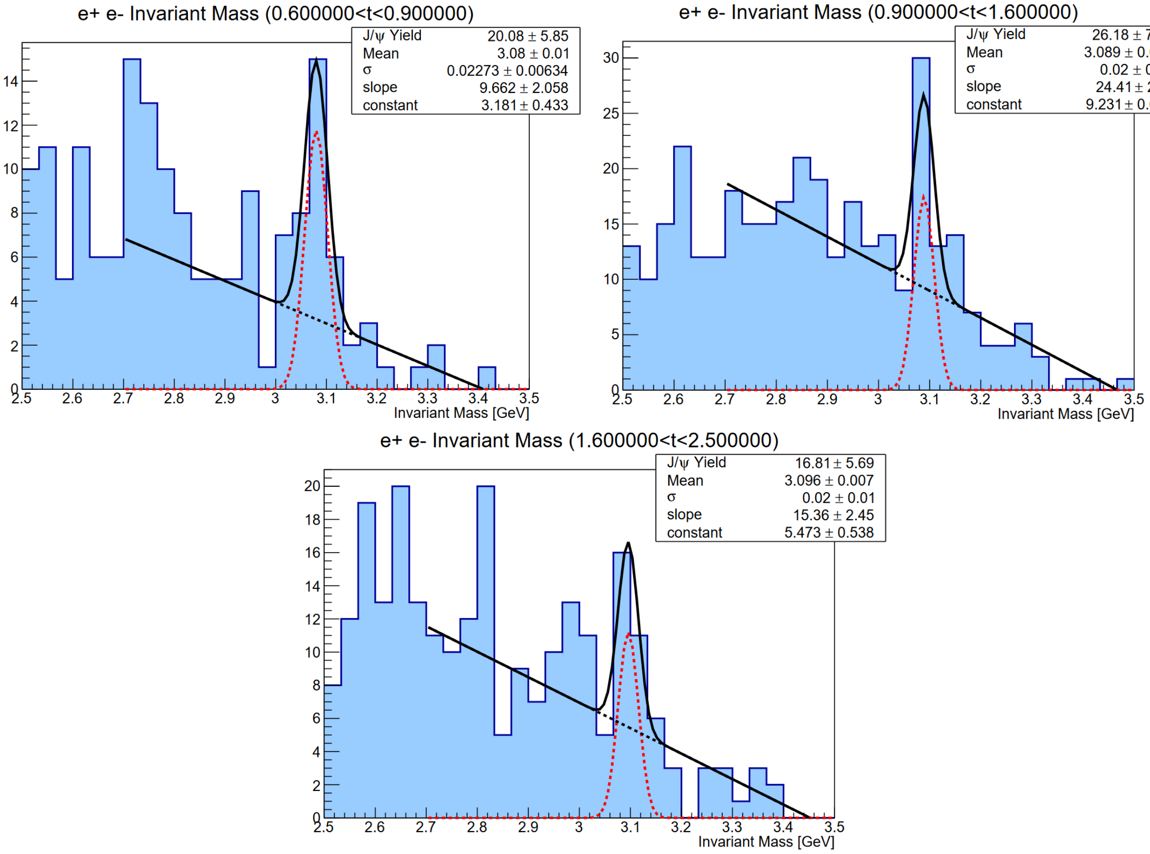


Figure B.8: The di-lepton invariant mass in the $en_{bound} \to (e')e^+e^-n$ channel when measuring the J/ψ differential cross section produced on the bound neutron.

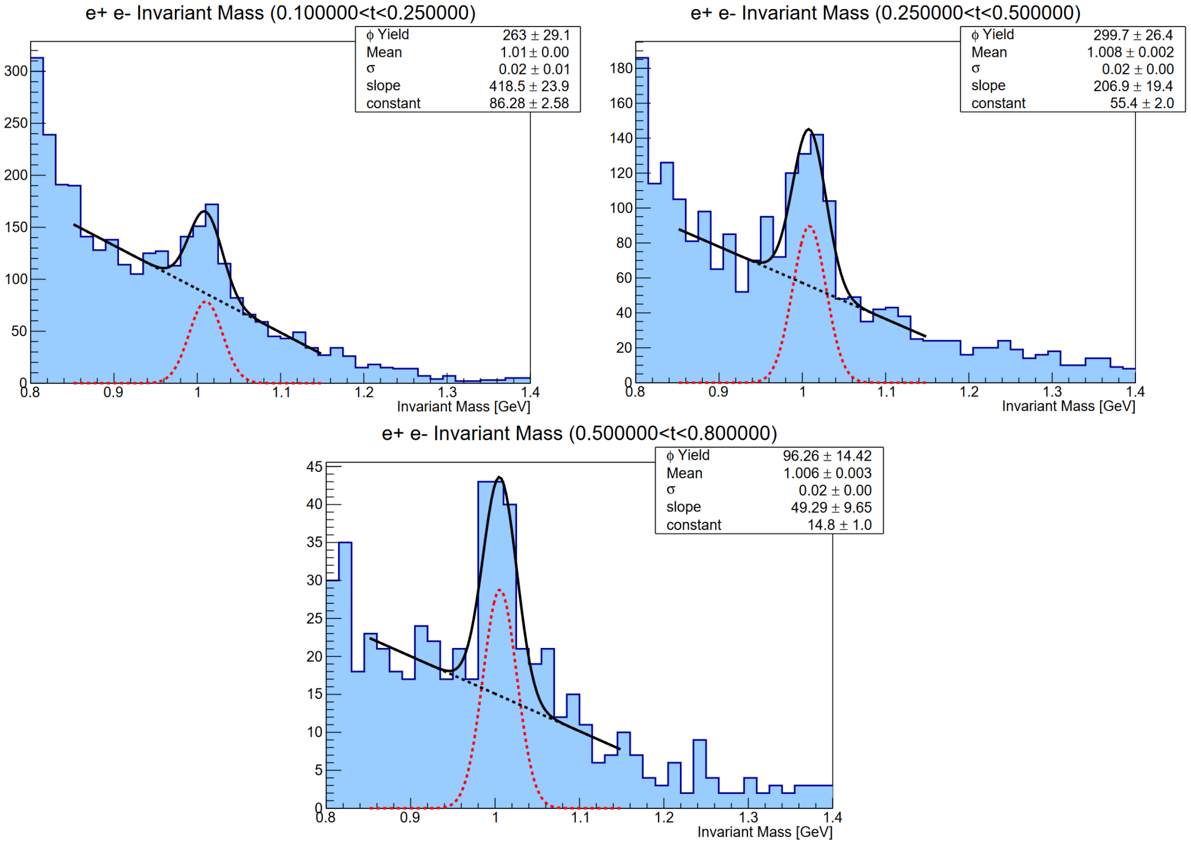


Figure B.9: The di-lepton invariant mass in the $ep \rightarrow (e')e^+e^-p$ channel when measuring the ϕ differential cross section produced on the free proton.

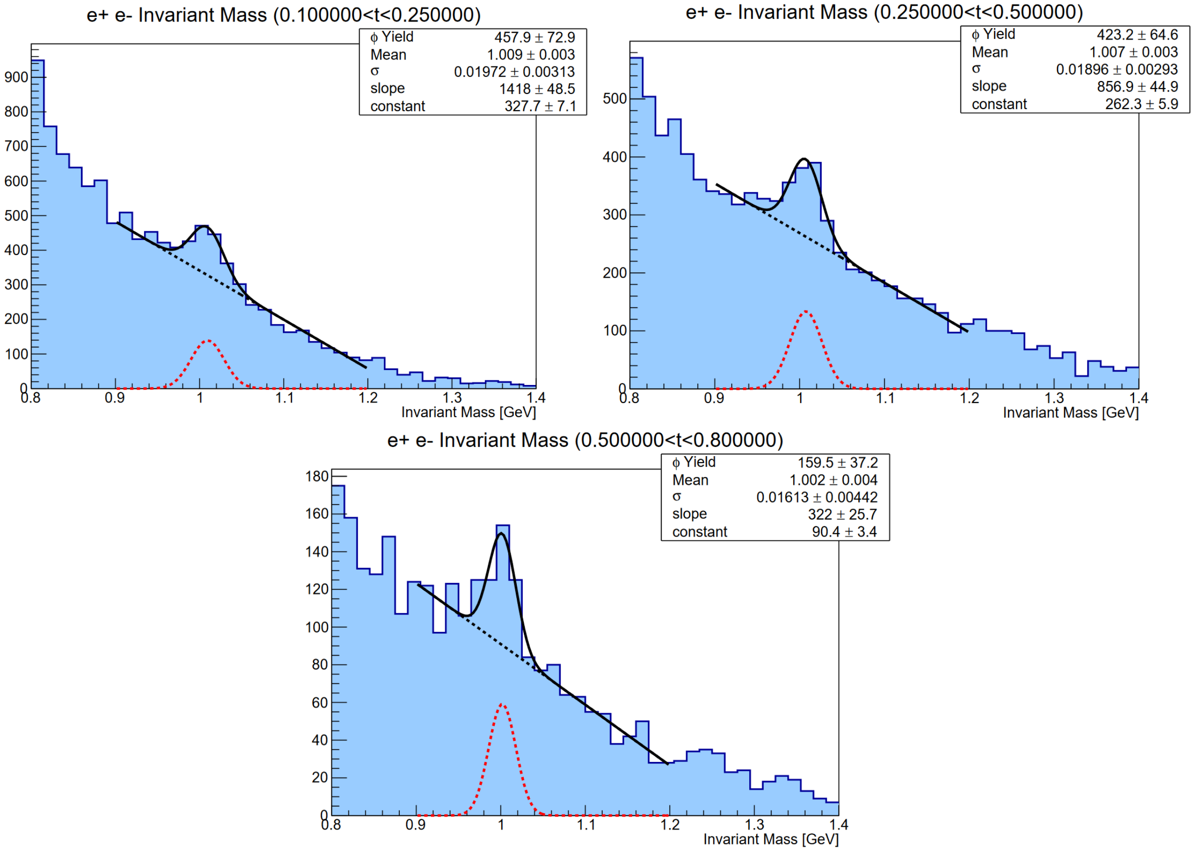


Figure B.10: The di-lepton invariant mass in the $ep_{bound} \to (e')e^+e^-p$ channel when measuring the ϕ differential cross section produced on the bound proton.

Bibliography

[1] J. Greensite, *An Introduction to the Confinement Problem*, Springer, 2011, Chapter 3

[2] D.J. Gross, F. Wilczek, Ultraviolet behavior of non-abelian gauge theories, *Phys. Rev. Lett.* **30** 26 1343–1346 (1973).

[3] H.D. Politzer, Reliable perturbative results for strong interactions, *Phys. Rev. Lett.* **30** 26 1346–1349 (1973).

[4] M. Thomson, *Modern particle physics*, Cambridge University Press, 2013, Chapter 5

[5] Wikipedia contributors, Standard Model, *Wikipedia, The Free Encyclopedia* (2023). Available online at https://en.wikipedia.org/wiki/Standard_Model (accessed 27 April 2023).

[6] R.L. Workman et al. (Particle Data Group), Review of Particle Physics, *Prog. Theor. Exp. Phys.* **2022** 083C01 (2022).
Available online at https://pdg.lbl.gov/ (accessed 10 August 2023).

[7] C. W. Misner, K.S. Thorne, J.A Wheeler, *Gravitation* W. H. Freeman 1973 Box 18.1.

[8] X.-D. Ji, QCD analysis of the mass structure of the nucleon, *Phys. Rev. Lett.* **74** 1071 (1995).

[9] J.J. Aubert, *et al.* (E598 Collaboration), Experimental Observation of a Heavy Particle J, *Phys. Rev. Lett.* **33** 1404 (1974).

[10] J.E. Augustin, *et al.* (SLAC-SP-017 Collaboration), Discovery of a Narrow Resonance in e+e− Annihilation, *Phys. Rev. Lett.* **33** 1406 (1974).

[11] J.D. Bjorken, The November Revolution: A theorist reminisces. Presented at the SLAC Symposium on the Tenth Anniversary of the November Revolution, *Fermilab-Conf-85/58* (1984).

[12] R.P. Crease, The November revolution, *Phys. World* **17** 18 (2004).

[13] H. Pagels, Energy-Momentum Structure Form Factors of Particles, *Phys. Rev.* **144** (1966).

[14] V.D. Burkert, L. Elouadrhiri, F.X. Girod, The pressure distribution inside the proton, *Nature* **557** 7705 (2018).

[15] V.D. Burkert, L. Elouadrhiri, F.Girod, Determination of shear forces inside the proton, *ArXiv e-prints* arXiv:2104.02031 (2021).

[16] Y. Hatta and D.-L. Yang, Holographic J/ψ production near threshold and the proton mass problem, *Phys. Rev. D* **98** 074003 (2018).

[17] C.J. Brown, D.H. Lyth, Assessment of the equivalent photon approximation for the process ee $\rightarrow$ ee $\pi + \pi -$, *Nucl. Phys. B*, **53** 2 (1973).

[18] T.H. Burnett, The diagonal equivalent photon density matrix: An exact extension of the equivalent photon approximation, *Nucl. Phys. B*, **114**, 1 (1976).

[19] S. Okubo, ϕ meson and unitary symmetry model, *Physics Letters* **5** 2 (1963).

[20] G. Zweig, *CERN Report No. 8419 / TH412* (1964).

[21] J. Iizuka, A Systematics and Phenomenology of Meson Family, *Progress of Theoretical Physics Supplement* **37** (1966).

[22] K. Tu (STAR Collaboration), J/psi photoproduction off deuteron in d+Au UPC at STAR (2020).
Presentation available online at https://indico.cern.ch/event/751767/contributions/3771014/attachme
v3.pdf (accessed 27 April 2023).

[23] D.J.Griffiths, *Introduction to elementary particles* Wiley, 2nd rev. version, 2008.

[24] J.J. Sakurai, Theory of strong interactions, *Annals of Physics* **11** 1 (1960).

[25] T.H. Bauer, R.D. Spital, D.R Yennie, F.M. Pipkin, The hadronic properties of the photon in high-energy interactions, *Reviews of Modern Physics* **50** 2 (1978).

[26] V. Barger, R.J.N. Phillips, Properties of ψN scattering, *Physics Letters B* **58** 4 (1975).

[27] S.J. Brodsky, E. Chudakov, P. Hoyer, J.M. Laget, Photoproduction of charm near threshold, *Phys.Lett. B* **498** 23 (2001).

[28] S.J. Brodsky, I. Schmidt, G.F. de Teramond, Nuclear-Bound Quarkonium, *Phys. Rev. Lett.* **64** 9 (1990).

[29] CLAS Collaboration, Near threshold J/ψ photoproduction and study of LHCb pentaquarks with CLAS12, Experiment Proposal (2017).
Available Online at https://www.jlab.org/exp_prog/proposals/17/E12-12-001A.pdf (accessed 27 April 2023).

[30] V. M. Braun, Higher Twists, *EPJ Web of Conference* **274** 01012 (2022).

[31] J.Erlich, An Introduction to Holographic QCD for Nonspecialists, *ArXiv e-prints* arXiv:1407.5002 (2014).

[32] T.-S. H. Lee, S. Sakinah, Yongseok Oh, Models of J/ψ photo-production reactions on the nucleon, *ArXiv e-prints* arXiv:2210.02154v1 (2022).

[33] R. Aaij *et al.* (LHCb Collaboration), Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\Psi K^- p$ Decays, *Phys. Rev. Lett.* **115** 072001 (2015).

[34] R. Aaij *et al.* (LHCb Collaboration), Observation of a narrow pentaquark state, $P_c(4312)^+$, and of two-peak structure of the $P_c(4450)^+$, *Phys.Rev.Lett.* **122** 22 (2019).

[35] A.N. Hiller Blin, C. Fernandez-Ramirez, A. Jackura, V. Mathieu, V.I. Mokeev, A. Pilloni, A. P. Szczepaniak, Studying the P_c (4450) resonance in J/ψ photoproduction off protons, *Few-Body Systems* **59** 104 (2018).

[36] M.I. Eides, V.Y. Petrov, M.V. Polyakov, New LHCb pentaquarks as hadrocharmonium states, *Mod.Phys.Lett.A* **35** 18 (2020).

[37] R. Chen, Z.-F. Sun, X. Liu, S.-L. Zhu, Strong LHCb evidence supporting the existence of the hidden-charm molecular pentaquarks, *Phys. Rev. D* **100**, 011502 (2019).

[38] F. Giannuzzi, Heavy pentaquark spectroscopy in the diquark model, *Phys. Rev. D* **99**, 094006 (2019).

[39] Wikipedia contributors, Pentaquark, *Wikipedia, The Free Encyclopedia* (2023).
Available online at https://en.wikipedia.org/wiki/Pentaquark (accessed 27 April 2023).

[40] U. Camerini, J.G. Learned, R. Prepost, C.M. Spencer, D.E. Wiser, W.W. Ash, R L. Anderson, D.M. Ritson, D.J. Sherden, C.K. Sinclair, Photoproduction of the ψ Particles, *Phys. Rev. Lett.* **35** 483 (1975).

[41] B. Gittelman, K. M. Hanson, D. Larson, E. Loh, A. Silverman, and G. Theodosiou, Photoproduction of the $\psi(3100)$ Meson at 11 GeV, *Phys. Rev. Lett.* **35** 1616 (1975).

[42] A. Ali *et al.* (GlueX Collaboration), First Measurement of Near-Threshold J/ψ Exclusive Photoproduction off the Proton, *Phys.Rev.Lett.* **123** 072001 (2019).

[43] A.N. Hiller Blin, C. Fernandez-Ramirez, A. Jackura, V. Mathieu, V.I. Mokeev, A. Pilloni, A. P. Szczepaniak, Studying the P_c (4450) resonance in J/ψ photoproduction off protons, *Phys. Rev. D* **94** 034002 (2016).

[44] D. Kharzeev, H. Satz, A. Syamtomov, and G. Zinovev, J/ψ-Photoproduction and the Gluon Structure of the Nucleon, *Nucl.Phys. A* **661** 568 (1999).

[45] M. I. Eides, V. Yu. Petrov, and M. V. Polyakov, Pentaquarks with Hidden Charm as Hadroquarkonia, *Eur. Phys. J. C* **78** 36 (2018).

[46] M. I. Eides and V. Yu. Petrov, Decays of pentaquarks in hadrocharmonium and molecular scenarios, *Phys. Rev. D* **98** 114037 (2018).

[47] F. Zeng, X.-Y. Wang, L. Zhang, Y.-P. Xie, R. Wang, and X. Chen, Near-threshold photoproduction of J/ψ in two-gluon exchange model, *Eur. Phys. J. C* **80** 1027 (2020).

[48] S. Adhikari, *et al.* (GlueX Collaboration), Measurement of the J/ψ photoproduction cross section over the full near-threshold kinematic region, *ArXiv e-prints* arXiv:2304.03845v1 (2023).

[49] L. Frankfurt, M. Strikman, Two-gluon form factor of the nucleon and J/ψ photoproduction, *Phys. Rev. D.* **66**, 031502 (2002).

[50] A. Martin, R. Roberts and W. Stirling, Structure Functions And Parton Distributions, *Int.J.Mod.Phys. A 10* 2885 (1995).

[51] D. Kharzeev and H. Satz, Quarkonium Interactions in Hadronic Matter, *Phys. Lett. B* **334** 155 (1994).

[52] S. Aid, *et al.*, Elastic and inelastic photoproduction of J/ψ mesons at HERA, *Nucl. Phys. B* **472** 3 (1996).

[53] D. Kharzaeev, Quarkonium Interactions in QCD, *Proc. Int. Sch. Phys. Fermi* **130** 105 (1996)
Preprint available on *ArXiv e-prints* arXiv:nucl-th/9601029 (1996).

[54] J.C. Collins, A. Duncan, S.D. Joglekar, Trace and dilatation anomalies in gauge theories, *Phys. Rev. D 16* 438 (1977).

[55] R. Wang, X. Chen, J. Evslin, The origin of proton mass from J/ψ photo-production data, *Eur. Phys. J. C* **80** 507 (2020).

[56] W. Kou, R. Wang, X. Chen, Extraction of Proton Trace Anomaly Energy from Near-Threshold ϕ and J/ψ photo-productions, *Eur. Phys. J. A* **58** 155 (2022).

[57] D. Kharzeev, Mass radius of the proton, *Phys. Rev. D* **104** 054015 (2021).

[58] D. Duran, *et al.* (J/ψ-007 Collaboration), Determining the gluonic gravitational form factors of the proton, *Nature* **615** 813–816 (2023).

[59] D. A. Pefkou, D. C. Hackett, P. E. Shanahan, Gluon gravitational structure of hadrons of different spin, *Phys. Rev. D* **105** 054509 (2022).

[60] P. Zyla, *et al.* (Particle Date Group), Review of Particle Physics, *Prog. Theor. Exp. Phys.* **083C01** (2020).

[61] Y.Guo, X.Ji, Y.Liu, QCD analysis of near-threshold photon-proton production of heavy quarkonium, *Phys. Rev. D* **103** 096010 (2021).

[62] R. Wang, W.Kou, Y.-P. Xie, X.Chen, Extraction of the proton mass radius from the vector meson photoproductions near thresholds, *Phys. Rev. D* **103** L091501 (2021).

[63] C. Han, G. Xie, W. Kou, R. Wang, X. Chen, The neutron and proton mass radii from the vector meson photoproduction data on the deuterium target, *Eur. Phys. J. A* **58** 105 (2022).

[64] R.L. Jaffe, Quark Distributions in Nuclei, *Phys. Rev. Lett.* **50** 4 (1983).

[65] J.J. Aubert *et al.* (European Muon Collaboration), The ratio of the nucleon structure functions F2N for iron and deuterium, *Phys. Lett. B.* **123B** 175 (1983).

[66] F.E. Close, R.G. Roberts, G.G. Ross, The effect of confinement size on nuclear structure functions, *Phys. Lett. B* **129** 5 (1983).

[67] G.A. Miller, Confinement in Nuclei and the Expanding Proton, *Phys. Rev. Lett.* **123** 232003 (2019).

[68] A.Donnachie, P.V.Landshoff, Gluon condensate and pomeron structure, *Nucl. Phys. B* **311** 3 (1989).

[69] E.Levin, Everything about reggeons. Part I:Reggeons in "soft" interaction, *ArXiv e-prints* arXiv:hep-ph/9710546 (1998).

[70] M.-L. Du, V. Baru, F.-K. Guo, C. Hanhart, U.-G. Meißner, A. Nefediev, I.Strakovsky, Deciphering the mechanism of near-threshold J/ψ photoproduction, *Eur. Phys. J. C* **80** 1053 (2020).

[71] I.I. Strakovsky, D. Epifanov, L. Pentchev, $J/\psi p$ scattering length from GlueX threshold measurements, *Phys. Rev. C* **101** 042201 (2020).

[72] L. Pentchev, I.I. Strakovsky, $J/\psi p$ scattering length from the total and differential photoproduction cross sections, *Eur. Phys. J. A* **57** 56 (2021).

[73] I.I. Strakovsky, L. Pentchev, A.I. Titov, Comparative analysis of ωp, ϕp, and $J/\psi p$ scattering lengths from A2, CLAS, and GlueX threshold measurements, *Phys. Rev. C* **101** 045201 (2020).

[74] S.J Brodsky, G.A. Miller, Is J / psi - nucleon scattering dominated by the gluonic van der Waals interaction?, *Phys.Lett.B* **412** 125 (1997).

[75] U. Skerbis, S. Prelovsek, Nucleon-J/ψ and nucleon-η_c scattering in P_c pentaquark channels from lattice QCD, *Phys. Rev. D* **99** 094505 (2019).

[76] A. Austregesilo (GlueX Collaboration), Spin-density matrix elements for vector meson photoproduction at GlueX, *AIP Conference Proceedings* **2249**, 030005 (2020).

[77] K. Schilling, P. Seyboth, G. Wolf, On the analysis of vector-meson production by polarized photons, *Nucl. Phys. B* **15** 2 (1970).

[78] L. Pentchev, New opportunities for J/ψ (and beyond) photoproduction studies in Hall-D with the CEBAF upgrade.
Presentation available online at https://indico.ectstar.eu/event/152/contributions/3133 /attachments/2001/2612/LPentchev_Jpsi_Trento.pdf (accessed 27 April 2023).

[79] J.P. Chen, H. Gao, T.K. Hemmick, Z.-E. Meziani, P.A. Souder, *et al* (SoLID Collaboration), A White Paper on SoLID (Solenoidal Large Intensity Device), *ArXiv e-prints* arXiv:1409.7741 (2014).

[80] Y.-Z. Xu, S.-Y. Chen, Z.-Q. Yao, D. Binosi, Z.-F. Cui, C.D. Roberts, Vector-meson production and vector meson dominance, *Eur. Phys. J. C* **81** 895 (2021).

[81] K. G. Boreskov, B. L. Ioffe, Photoproduction of J/psi mesons in the peripheral model. The diffraction mechanism, *Sov. J. Nucl. Phys.* **25** 331 (1977).

[82] CLAS Collaboration, Jefferson Lab PAC 39 Proposal Timelike Compton Scattering and J/ψ photoproduction on the proton in e^+e^- pair production with CLAS12 at 11 GeV, Experiment Proposal (2012).
Available online at https://www.jlab.org/exp_prog/proposals/12/PR12-12-001.pdf (accessed 27 April 2023).

[83] M.D. Baker, A.J. Freese, L. Guo, C. Hyde, Y. Ilieva, B. McKinnon, P. Nadel-Turonski, M. Sargsian, V. Kubarovsky, S. Stepanyan, N. Zachariou, Z.W. Zhao, Study of J/ψ Photoproduction off Deuteron, Experiment Proposal (2018).
Available Online at https://www.jlab.org/exp_prog/proposals/18/E12-11-003B.pdf (accessed 27 April 2023).

[84] J. Newton, J/ψ Photoproduction Near Threshold With CLAS12, Ph.D. Thesis Old Dominion University (2021). Available online at https://digitalcommons.odu.edu/physics_etds/134/ (accessed 27 April 2023).

[85] R.D. McKeown, The Jefferson Lab 12 GeV Upgrade, *J. Phys.: Conf. Ser.* **312** 032014 (2011).

[86] H.A. Grunder, J.J. Bisognano, W.I. Diamond, B.K. Hartline, C.W. Leemann, J. Mougey, R.M. Sundelin, R.C. York, The continuous electron beam accelerator facility, *Nuclear Physics A* **478** (1988).

[87] C. Tennant, A. Carpenter, K. Iftekharuddin, T. Powers, M. Rahman, Superconducting Radio-Frequency Cavity Fault Classification Using Machine Learning at Jefferson Laboratory, *Phys. Rev. Accel. Beams* **23** 114601 (2020)

[88] V.D. Burkert, *et al.*, The CLAS12 Spectrometer at Jefferson Laboratory, *Nucl. Inst. Methods A* **959** 163419 (2020).

[89] A. Celentano, The Forward Tagger detector for CLAS12 at Jefferson Laboratory and the MesonEx experiment, Ph.D. Thesis Università degli Studi di Genova (2014). Available online at https://www.ge.infn.it/jlab12/files/tesi_2014_03_celentano_andrea.pdf (accessed 27 April 2023).

[90] The CLAS Collaboration, CLAS12 Technical Design Report V5.1 (2008), available online at https://www.jlab.org/Hall-B/clas12_tdr.pdf (accessed 27 April 2023).

[91] A. Acker, *et al.*, The CLAS12 Forward Tagger, *Nucl. Inst. Methods A* **959** 163475 (2020).

[92] Y.G. Sharabian, *et al.*, The CLAS12 high threshold Cherenkov counter, *Nucl. Inst. Methods A* **968** 163824 (2020).

[93] M. Ungaro, *et al.*, The CLAS12 Low Threshold Cherenkov Counter, *Nuclear Inst. Methods A* **957** 163420 (2020).

[94] M.D. Mestayer, *et al.*, The CLAS12 drift chamber system, *Nucl. Inst. Methods A* **959** 163518 (2020)

[95] D.S. Carman, *et al.*, The CLAS12 Forward Time-of-Flight system, *Nuclear Inst. Methods A* **960** 163629 (2020).

[96] G. Asryan, *et al.*, The CLAS12 forward electromagnetic calorimeter, *Nucl. Inst. Methods A* **959** 163425 (2020).

[97] V.D. Burkert, Overview of CLAS12, Workshop on RICH Detector for CLAS12. Presentation available online at https://www.jlab.org/conferences/CLAS12/Burkert.pdf (accessed 27 April 2023).

[98] A. Acker, *et al.*, The CLAS12 Micromegas Vertex Tracker, *Nucl. Inst. Methods A* **957** 163423 (2020).

[99] M.A. Antonioli, *et al.*, The CLAS12 Silicon Vertex Tracker, *Nucl. Inst. Methods A* **962** 163701 (2020).

[100] D.S. Carman, *et al.*, The CLAS12 Central Time-of-Flight system, *Nucl. Inst. Methods A* **960** 163626 (2020).

[101] P.Chatagnon, *et al.*, The CLAS12 Central Neutron Detector, *Nucl. Inst. Methods A* **959** 163441 (2020).

[102] S. Niccolai, Neutron detector for the central part of CLAS12 (2008). Presentation available online at https://slideplayer.com/slide/13267096/ (accessed 27 April 2023).

[103] V. Ziegler *et al.*, The CLAS12 software framework and event reconstruction, *Nucl. Inst. Methods A* **959** 163472 (2020).

[104] E. Segre, H. Staub, H. Bethe, J. Ashkin, Experimental Nuclear Physics, John Wiley and sons (1953).

[105] P. Sempere Roldan, Quality control and preparation of the PWO crystals for the electromagnetic calorimeter of CMS, Ph.D. Thesis Universidade de Santiago de Compostela (2011).
Available online at https://cds.cern.ch/record/1388922 (accessed 27 April 2023).

[106] P. Thomadakis, A. Angelopoulos, G. Gavalian, N. Chrisochoides, De-noising drift chambers in CLAS12 using convolutional auto encoders, *Computer Physics Communications* **271** 108201 (2022).

[107] G. Gavalian, P. Thomadakis, A. Angelopoulos, N. Chrisochoides, R. De Vita, V. Ziegler, CLAS12 Track Reconstruction with Artificial Intelligence, *ArXiv e-prints* arXiv:2202.06869 (2022).

[108] G. Gavalian, Artificial Intelligence in CLAS12, CLAS Collaboration Meeting. Presentation available online at
https://indico.jlab.org/event/536/contributions/10120/attachments/7970/11249/CLAS-Collaboration-June-24-2022.pdf (accessed 27 April 2023).

[109] B. Raydo, *et al.*, The CLAS12 Trigger System, *Nucl. Instrum. Methods A* **960** 163529 (2020).

[110] G. Gavalian, *HIPO*, https://github.com/gavalian/hipo (accessed 27 April 2023).

[111] D. Glazier, *clas12root*, https://github.com/JeffersonLab/clas12root (accessed 27 April 2023).

[112] D. Glazier, *chanser*, https://github.com/dglazier/chanser (accessed 27 April 2023).

[113] S. Agostinelli et al., Geant4 - A Simulation Toolkit, *Nucl. Instrum. Meth. A* **506** 250303 (2003).

[114] M. Ungaro *et al.*, The CLAS12 Geant4 simulation, *Nucl. Instrum. Methods A* **959** 163422 (2020).

[115] D. Glazier, *elSpectro*, https://github.com/dglazier/elSpectro/tree/master (accessed 27 April 2023).

[116] Kim Albertsson *et al.*, Machine Learning in High Energy Physics Community White Paper, *J. Phys.: Conf. Ser.* **1085** 022008 (2018).

[117] A. Burkov, *The Hundred-Page Machine Learning Book* 2019 Open Access at http://themlbook.com/wiki/doku.php (accessed 27 April 2023).

[118] V. Pleskot, D. Scheirich, Statistical Methods (2020).
YouTube channel available at https://www.youtube.com/@statisticalmethods6041 (accessed 27 April 2023).

[119] G. Sanderson, 3Blue1Brown (2015)
YouTube channel available at https://www.youtube.com/c/3blue1brown (accessed 27 April 2023).

[120] R. Tyson, https://github.com/rtysonCLAS12/ML_notebooks/blob/main/Toy_TrackDenoiser.ipynb (accessed 11 August 2023).

[121] Unbabel Inc, Why AI fails in the wild, *Unbabel* (2019).
Available online at https://resources.unbabel.com/blog/artificial-intelligence-fails (accessed 27 April 2023).

[122] R. Tyson, G. Gavalian, D. Ireland, B. McKinnon, Deep Learning Level-3 Electron Trigger for CLAS12, *ArXiv e-prints* arXiv:2302.07635 (2023).

[123] M.Cavaioni, Machine Learning: Decision Tree Classifier, *Medium* (2017).
Available online at https://medium.com/machine-learning-bites/machine-learning-decision-tree-classifier-9eb67cad263e (accessed 27 April 2023).

[124] J. Furnkranz, *Encyclopedia of Machine Learning*, Springer US, 2010, pp 263–267.

[125] S. K. Murthy, Automatic construction of decision trees from data: A multi-disciplinary survey, *Data Mining and Knowledge Discovery* **2** 4 (1998).

[126] Y. Freund, R. E. Schapire, Experiments with a new boosting algorithm, *Proceedings of the 13th international conference on machine learning* 148–156 (1996).

[127] J.H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of Statistics* **29** 5 (2001).

[128] A. Connolly (AstroML developers), Deep Learning: Classifying Astronomical Images, *AstroML Interactive Book* (2022).
Available online at https://www.astroml.org/astroML-notebooks/chapter9/astroml_chapter9_Deep_Learning_Classifying_Astronomical_Images.html (accessed 27 April 2023).

[129] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.

[130] D. P. Kingma and J. Ba, Adam, Adam: A method for stochastic optimization, *ArXiv e-prints* arXiv:1412.6980 (2014).

[131] F. Murtagh, Multilayer perceptrons for classification and regression, *Neurocomputing*, **2** 5 (1991).

[132] S. Albawi, T. A. Mohammed and S. Al-Zawi, Understanding of a convolutional neural network, *2017 International Conference on Engineering and Technology (ICET)*, doi: 10.1109/ICEngTechnol.2017.8308186, (2017).

[133] K. O'Shea, and R. Nash, An Introduction to Convolutional Neural Networks, *ArXiv e-prints* arXiv:1511.08458 (2015).

[134] D. Bank, N. Koenigstein, R. Giryes, Autoencoders, *ArXiv e-prints* arXiv:2003.05991 (2020).

[135] M. Abadi, *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, *ArXiv e-prints* arXiv:1603.04467v2 (2016). Software available from tensorflow.org.

[136] F. Pedregosa, *et. al.*, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **12** 2825 (2011).

[137] Eclipse Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0. https://deeplearning4j.konduit.ai/ (accessed 27 April 2023).

[138] J. Therhaag, TMVA Toolkit for multivariate data analysis in ROOT, *PoS* **ICHEP2010** 510 (2010).

[139] K. Albertsson, *et. al.*, ROOT TMVA user guide, available online at https://root.cern.ch/download/doc/tmva/TMVAUsersGuide.pdf (accessed 27 April 2023).

[140] H. Avakian, *et al.*, CLAS12 RG-A - Analysis Note Overview and Procedures Phase I Towards SIDIS CLAS12 First Publications, CLAS12 Analysis Note (2020).
Available online at https://www.jlab.org/Hall-B/shifts/admin/paper_reviews /2020/RGA_Analysis_Overview_and_Procedures_Nov_4_2020-6245173-2020-12-09-v3.pdf (accessed 27 April 2023).

[141] P. Chatagnon, *et al.* (CLAS Collaboration), First Measurement of Timelike Compton Scattering, *Phys. Rev. Lett.* **127**262501 (2021).

[142] P. Chatagnon, Nucleon structure studies with CLAS12 at Jefferson Lab : timelike Compton scattering and the central neutron detector, Ph.D. Thesis Université Paris-Saclay (2020).
Available online at https://www.theses.fr/2020UPASP039 (accessed 27 April 2023).

[143] S. Stepanyan, Detection of neutrons using the g10 data set, CLAS Analysis Note (2008).
Available online at https://userweb.jlab.org/∼stepanya/g10/g10_n_detection.pdf (accessed 27 April 2023).

[144] R. Paremuzyan, Timelike Compton Scattering, Ph.D. Thesis Yerevan Physics Institute (2010).
Available online at https://www.jlab.org/Hall-B/general/thesis/Paremuzyan_thesis.pdf (accessed 27 April 2023).

[145] P. Thomadakis, K. Garner, G. Gavalian, N. Chrisochoides, Charged particle reconstruction in CLAS12 using Machine Learning, *Computer Physics Communications* **287** 108694 (2023).

[146] P. Chatagnon, J. Newton, Private Communication.

[147] D. Glazier, brufit, https://github.com/dglazier/brufit/tree/dev/tutorials (accessed 27 April 2023).

[148] M. Pivk, F.R. Le Diberders, sPlot: a statistical tool to unfold data distributions, *Nucl. Instrum. Methods A*, **555** 1 (2005).

[149] M. Pivk, sPlot: A Quick Introduction, *ArXiv e-prints* arXiv:physics/0602023v1 (2006).

[150] S. Frixione, M.L. Mangano, P. Naso and G. Ridolfi, Improving the Weizsacker-Williams approximation in electron-proton collisions, *Phys. Lett. B* **319** (1993).

[151] P. Hurck, Private Communication.

[152] R. Barlow, Systematic errors: Facts and fictions, *Conference on Advanced Statistical Techniques in Particle Physics* (2002). Preprint available on *ArXiv e-prints* arXiv:hep-ex/0207026.

[153] CLAS12 Collaboration, *RCDB*, https://github.com/JeffersonLab/rcdb and https://clasweb.jlab.org/rcdb/ (accessed 27 April 2023).

[154] C. Dilks, *QADB*, https://github.com/JeffersonLab/clasqaDB/tree/master (accessed 27 April 2023).

[155] D. Darulis, R. Tyson, D. G. Ireland, D. I. Glazier, B. McKinnon, P. Pauli, Machine Learned Particle Detector Simulations, *ArXiv e-prints* arXiv:2207.11254 (2022).

List of recurrent acronyms:

- AI: Artificial Intelligence (see Chapter 5).
- BDT: Boosted Decision Tree (see Chapter 5).
- CD: CLAS12 Central Detector (see Chapter 3).
- CEBAF: Continuous Electron Beam Accelerator Facility (see Chapter 3).
- CED: CLAS12 Event Display (see Chapter 3).
- CLAS12: CEBAF Large Acceptance Spectrometer (see Chapter 3).
- clas12root: Analysis software for CLAS12 (see Chapter A.2).
- chanser: Analysis software for CLAS12 (see Chapter A.2).
- CNN: Convolutional Neural Network (see Chapter 5).
- DC: CLAS12 Forward Drift Chambers (see Chapter 3).
- DT: Decision Tree (see Chapter 5).
- DVCS: Deeply Virtual Compton Scattering (see Chapter 1).
- EB: Event Builder (see Chapter 3).
- ECAL: CLAS12 Forward Electromagnetic Calorimeter, composed of the preshower calorimeter PCAL and the electromagnetic calorimeter EC, itself made of the ECIN and ECOUT (see Chapter 3).
- EMC Effect: European Muon Collaboration effect (see Chapter 2).
- EMT: Energy Momentum Tensor (see Chapter 2).
- EPA: Equivalent Photon Approximation (see Chapter 2).
- FD: CLAS12 Forward Detector (see Chapter 3).
- FT: CLAS12 Forward Tagger (see Chapter 3).

- FTOF: CLAS12 Forward Time of Flight detector (see Chapter 3).
- GFF: Gravitational Form Factor (see Chapter 2).
- HIPO: High Performance Output data format (see Chapter 3).
- HTCC: CLAS12 Forward High Threshold Cherenkov Counter (see Chapter 3).
- JLAB: Thomas Jefferson National Accelerator Facility (see Chapter 1).
- LTCC: CLAS12 Forward Low Threshold Cherenkov Counter (see Chapter 3).
- MIP(s): Minimum Ionising Particle(s) (see Chapter 6).
- ML: Machine Learning (see Chapter 5).
- MLP: Multilayer Perceptron (see Chapter 5).
- NN: Neural Network (see Chapter 5).
- PID: Particle Identification (see Chapter 6).
- PMT: Photomultiplier Tubes (see Chapter 3).
- QCD: Quantum Chromodynamics (see Chapter 2).
- RID: Reaction Identification (see Chapter 7).
- SM: Standard Model (see Chapter 1).
- SDME(s): Spin Density Matrix Element(s) (see Chapter 2).
- TCS: Timelike Compton Scattering (see Chapter 7).
- TMVA: ROOT Toolkit for Multivariate Analysis (see Chapter 5).
- U/V/W: Views in the ECAL (see Chapter 3).
- VMD: Vector Meson Dominance (see Chapter 2).