CHAPTER 5

DATA ANALYSIS

The first production run at JLab's Hall B after the completion of the 12 GeV upgrade was called Run Group A. It ran over the fall of 2018 and spring 2019 with billions of events and 2 pB of data accumulated. The beam energy for the run was 10.6 GeV on a 5 cm liquid hydrogen target at a current between 5 nA to 75 nA. Because this is the first set of data coming from CLAS12 detectors, it serves as the data from which all important calibration and analysis is done to better understand all subsequent run groups.

This chapter will outline the software tools used for turning raw data into usable data sets, and the calibrations done to correct the raw data. Then we will begin the analysis portion, extracting the inclusive DIS cross section from Run Group A. Here we must make fiducial cuts to eliminate inefficient areas of individual detectors and kinematic cuts to select DIS events. Finally we put the resulting data together into bins of x and y (where $y = \nu/E$) to find the acceptance and calculate the inclusive DIS cross section. The purpose of extracting this cross section is to compare it the well-established Christy-Bosted fits to data.

5.1 CLAS12 OFFLINE SOFTWARE

The raw data coming from each detector first enters into the ReadOut Controller (ROC) and then gets stored in the EVent Input Output (EVIO) format. EVIO is a data format that is designed and maintained by JLab Data Acquisition Group. Once that data is available for off-line use, it requires decoding. Decoding is the process of taking EVIO raw data and converting it to High Output Performance Output (HIPO) format.

The HIPO format provides for a flexible data container structure, and minimizes disk space by utilizing LZ4 data compression (the fastest compression method currently available). In each HIPO file, data is stored as individual records with adjustable size. Each record is compressed, with a tag associated with it, and a pointer to it is stored in the file's index table. For analysis this provides users with faster analysis by reading portions of the file depending on the final states to be analyzed.

Once the data is in HIPO format, it is ready to by reconstructed and analyzed. The CLAS12 event Reconstruction and Analysis (or CLARA) framework allows users to reconstruct physics events and analyze the files to yield usable physical data. CLARA does this by utilizing an service-oriented architecture to enhance agility, efficiency, and productivity of the software components within the CLARA framework. During the reconstruction process, the raw data from all detectors is taken in and processed by the corresponding packages. The main packages in CLARA are for *geometry*, *calibration constants*, *magnetic fields*, *particle swimming*, and *plotting/-analysis*.

The geometry tools were created due to the complexity of the CLAS12 subsystem geometries. The library contains primitives that represent all of the lines, planes and shapes of all the detectors. The tools provide methods to track particles through the different volumes for evaluation of track trajectories, such as line-to-surface intersections, ray tracing through objects, and evaluation of the distance of closest approach to a line or surface. Because subsystem parameters can change from run group to run group and sometimes even within a run group, time-dependent geometry variations exist that allows for consistency between simulation, reconstruction, and event visualization packages.

The Calibration Constants Database (CCDB) was originally developed at JLab for the GlueX Experiment in Hall D. It was adopted by CLAS12 group because of its functionality for storing and accessing structured tables. At the decoding stage, file formats change, but also data structures. Signals are converted from hardware notation (*i.e.* crate, slot, channel) to CLAS12 notation (*i.e.* sector, layer, component). Then during reconstruction, the time stamps of these databases are utilized in order to access run-specific constants. The CLAS12 software tools employ an Application Programming Interface (API) that parses CCDB tables to create structured maps of the constants stored in memory by sector, layer, component. This method allows for fast retrieval of only the relevant constants.

Magfield, the magnetic field package for CLARA, consists of binary field maps created from engineering models of the solenoid and torus magnets in CLAS12. These field maps contain a header with meta-data describing field pedigree, its grid coordinate system, and the coordinate system of the field components. Because the field is often accessed within a sequence of points all contained within a single grid, *magfield* uses time-saving software probes to cache nearest neighbors.

To propagate charged particles through the CLAS12 magnetic fields, the *swimmer* package, in parallel with the *magfield* package, is used. Swimmer uses a fourth-order adaptive-size Runge-Kutta integrator with single step advancement achieved by a configurable Butcher tableau advancer. The purpose of swimming particles with this toolkit is to propagate particles to a given plane, to the closest point on a line, or to a given (x, y, z) coordinate. Performance is improved for forward propagation in CLARA by reducing the dimensionality of a state vector that contains the main track parameters, by changing from the path length independent variable to the coordinate along the beamline, which defines the nominal CLAS12 z-axis.

Finally, the plotting and analysis tools can be used for further data calibration, monitoring, and analysis. The toolkit was developed in the Java programming language and the interface is similar the ROOT platform developed at CERN for high-energy physics analysis. The plotting package, called *groot*, allows for histogram and graph creation, filling and manipulation. Plot fitting can be done using the Java-based MINUIT library available in the JHEP repositories.

Once the information about particle tracks is collected, that information is passed to a service called the Event Builder (EB). The EB takes the results from the upstream services and correlates the information from the CLAS12 subsystems. To form charged particles from the data, EB matches geometric coincidences in the distance of closest approach (DOCA) between detector responses and tracks. The event start time is important for all time-based particle identification and is determined from the optimal charged particle candidate in the Forward Detectors with an associated FTOF timing response. The last step in the EB is particle identification. For our purposes, we are really only concerned with e^- identification. This e^- PID is largely done through calorimetry and Cherenkov information. If the measured energy deposition in the ECAL is consistent within 5σ of the expected value of the sampling fraction, and the photoelectron response in the HTCC is consistent with $\beta \approx 1$, then the particle is assigned to be an electron or positron depending on the track curvature in the DC.

5.2 CALIBRATION

Once the raw data is decoded and reconstructed, it can be analyzed. However, initial analysis must be dedicated to detector calibration. Calibration is done for each detector and even for each run so that the experimental quantities like time and

energy are correctly extracted from raw TDC and ADC data. Just as in the RTPC, drift times and distances of electrons in the DC are subject to the properties of the gas (*i.e.* pressure, temperature, gas mixture, etc.). These changes determine calibration constants for the DC, just as they do for the RTPC. TOF calibration constants change with changes in the wires or electronics. The calibrations of individual detectors have been done by a large group of CLAS12 collaborators. Those calibration efforts will be briefly discussed and focused on the detectors relevant to this analysis.

The order of calibrating the detectors was important since some calibrations rely on the proper calibrations of other detectors. The first step was the DC calibration with FTOF time matching. This relied on a crude start time (few ns level) calibration of the FTOF, whose offset requires calibration between the FADC and TDC. Then the data needed to be recooked, which means that it required a run through CLARA again for reconstruction given the new calibration constants. After the recooking, the FTOF was calibrated more precisely with CTOF time matching. Energy calibrations for the FTOF could be done before the DC calibration using crude DC calibration parameters for path length corrections, but ideally done post DC calibration. FTOF timing calibrations employed PID from the Event Builder, and defined the start time using the electron in the EC, positron in the EC, or high-momentum pion in the DC/FTOF. Another recooking was necessary to implement the new calibration constants from CCDB.

Once the DC and FTOF were properly calibrated, CLAS12 subsystem were calibrated. This included CND, CTOF, EC, FT (hodoscope and calorimeter), HTCC, LTCC, and RICH. Timing calibrations for all subsystems relied on PID from the EB and start time from the FTOF. The energy calibrations for the subsystems only employed PID from the EB. Recooking was again necessary after subsystem calibration to update the reconstructed data using the new CCDB parameters. Lastly, the RF calibration was done to capture the overall RF time shifts run by run.

Run	Torus	Solenoid	$< i >$ [nA]	E_{beam} [GeV]	Run Range
4903	-100%	-100%	45	10.6	4763-5031
5038	-100%	-100%	45	10.6	5032-5189
5197	-100%	-100%	45	10.6	5190-5285
5306	-100%	-100%	45	10.6	5286-5419

TABLE III: Summary of calibrations done for Run Group A.

The resulting calibrations can be summarized in Table III. The required specifications for calibration were generally met. For the DC, a requirement that δx = 250-400 µm was not met since after calibration δx = 330-400 µm. However, for FTOF, δt = 60-110 ps (p1b) and after calibration δt = 60-120 ps (p1b). For the EC, a requirement that $\sigma_E/E = 10\%/\sqrt{E}$ was met exactly after calibration and the $< t_\gamma > <$ 500 ps was also met.

5.3 FIDUCIAL CUTS

Each detector has limits where it cannot efficiently detect particles. Near the edges of detectors are particularly vulnerable to inefficient particle detection. The goal of placing fiducial cuts on detectors is to minimize ineffective areas of each detector while maximizing the number of "good" particles we keep.

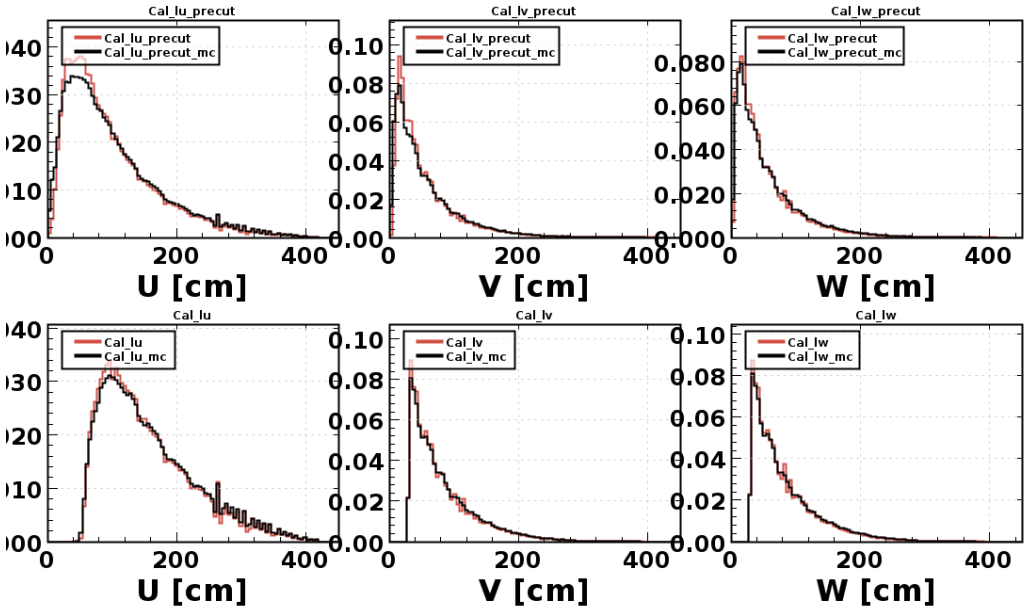

Fig. 35: 1D distributions of the U, V, and W sectors of the PCAL. The uncut histograms are on the top and the bottom histograms contain the cuts: $U < 30$ cm, $30 < V < 390$ cm, and $30 < W < 390$ cm. All plots are normalized to account for any mismatch in total statistics. Red lines are from RGA data and black lines are for the Monte-Carlo (MC) data.

The EC is the detector that we use for determining the electron four-momentum and all kinematics that are calculated from that momentum. When electrons enter the EC they shower and stop. That EM shower is broad, so we have to remove events

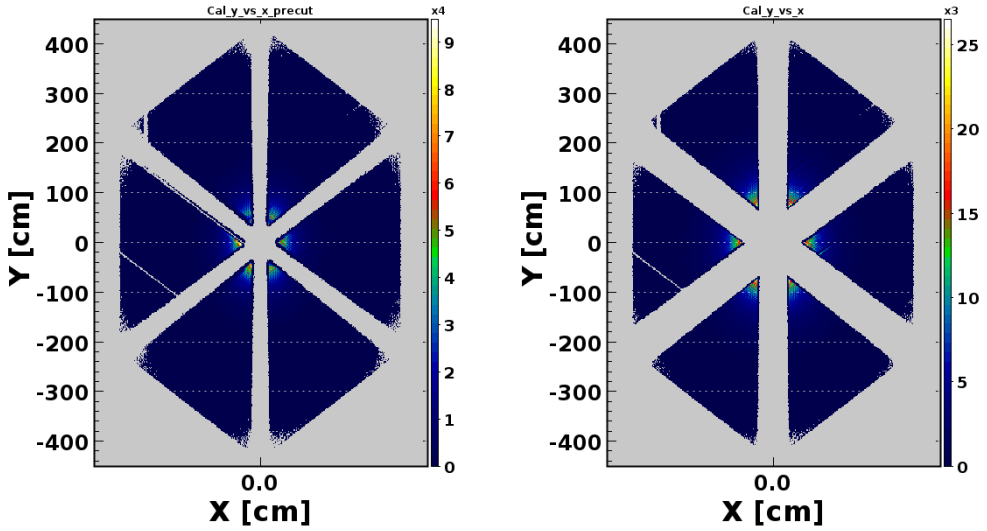

Fig. 36: 2D histograms of the uncut PCAL hits on the left and the 2D histogram containing the fiducial cuts.

close to the edges as showers leak out and the sampling fraction becomes unreliable for e^- PID.

Remember that the EC sector is defined by U, V, and W edges of its triangular shape. We can cut on those edges for the PCAL only and the effects will propagate through to the EC_{inner} (ECin) and EC_{outer} (ECout). The established cuts for the PCAL are $U < 30$ cm, $30 < V < 390$ cm, and $30 < W < 390$ cm. Fig. 35 shows the one-dimensional distributions of the U, V, and W sectors of the PCAL. The uncut histograms are on the top and the bottom histograms contain the cuts that were described. Fig. 36 contains the two-dimensional histograms of the uncut PCAL hits on the left and the 2D histogram containing the fiducial cuts.

5.4 KINEMATIC CUTS

Our goal is to extract the inclusive DIS cross section for the process $ep \longrightarrow e'X$, which means that certain constraints must be put on some of the kinematic variables. To isolate DIS events, $W > 2$ GeV and $Q^2 > 1$ GeV^2. Other *kinematic cuts* must be applied. In order to isolate events that originate at the target, we require $-10\text{cm} < v_z < 10\text{cm}$, where v_z is the z-vertex position of the track. Fig. 37 shows v_z before (left) and after (right) cuts. Because there is no discernible difference between left and right (uncut and cut), it is clear that cuts were made during reconstruction.

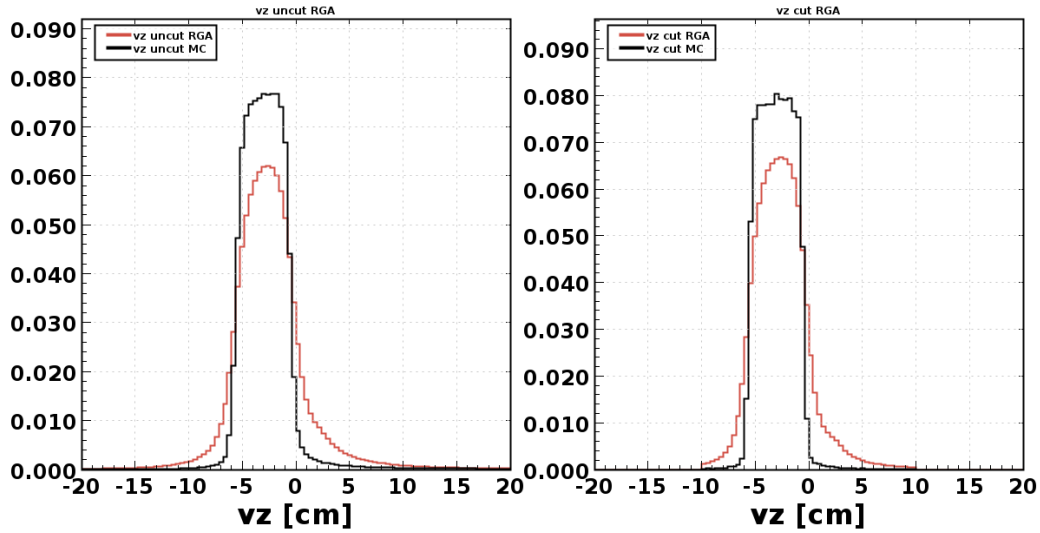

Fig. 37: v_z before and after all described cuts. Red lines are from RGA data and black lines are for the Monte-Carlo (MC) data.

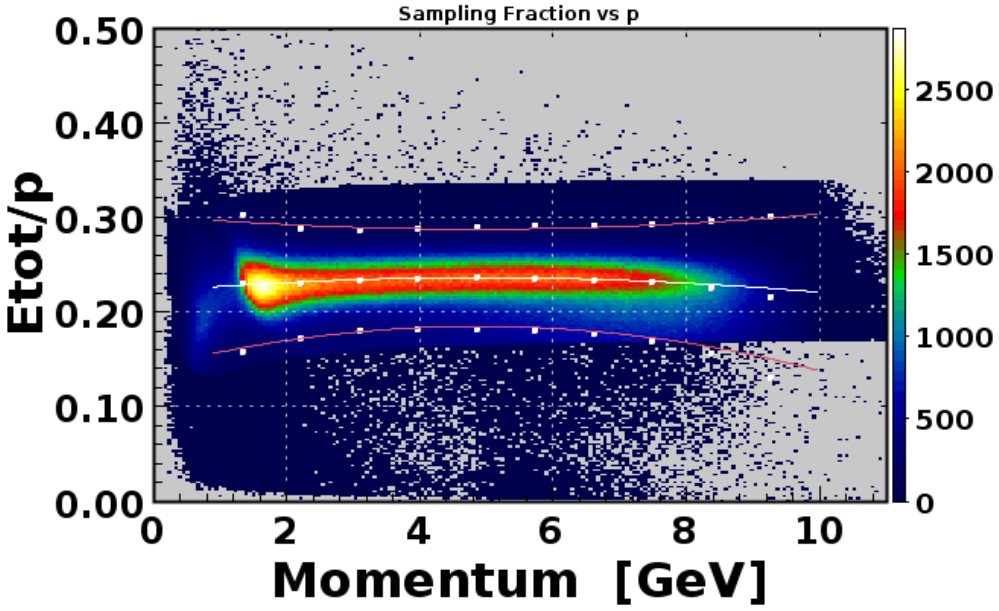

Fig. 38: Slicing and fitting of the sampling fraction to create a 5σ cut.

The last significant kinematic cut occurs on the sampling faction. The sampling fraction is defined as E_{tot}/p, where E_{tot} is the total energy deposited by the particle in all three layers of the EC and p is the particle's momentum. When the sampling

fraction is plotted vs p, electrons will appear as a band around 0.25 and with a momentum of above 1 GeV to eliminate any minimum ionizing particles like pions. To do this, we take slices of the plot, find the mean and sigma of E_{tot}/p for that slice and cut out anything $\pm 2.5\sigma$. We do this for values along p and fit the points for means and $\pm 2.5\sigma$ to polynomials that we can cut on to isolate electrons.

Because of the Forward Detector's coverage in θ, we need to make sure that $5^\circ < \theta < 40^\circ$ so it falls within that coverage. Fig. 39 shows the uncut kinematic variables and Fig. 40 was created with the kinematic cuts described. Fig. 41 and Fig. 42 shows uncut and cut (respectively) energy distributions for E', EC, and the number of photoelectrons in the HTCC (*nphe*). The next group of plots Fig. 43-48 shows two-dimensional histograms of various kinematic variables all uncut and after applying all cuts. In Fig. 48 one can clearly see the successful application of the 5σ cut described in the previous paragraph.

The summary of applied cuts is as follows:

- $W > 2$ GeV
- $Q^2 > 1$ GeV^2
- $5^\circ < \theta < 40^\circ$
- PCAL fiducial cuts: $U < 30$ cm, $30 < V < 390$ cm, and $30 < W < 390$ cm
- $-10 < v_z < 10$ cm
- 5σ cut on sampling fraction
- HTCC cut: $nphe > 5$
- EC energy cuts: $E_{\mathrm{PCAL}} > 0.06$ GeV, $E_{\mathrm{EC_{in}}} > 0.025$ GeV, $E_{\mathrm{EC_{out}}} > 0.05$ GeV

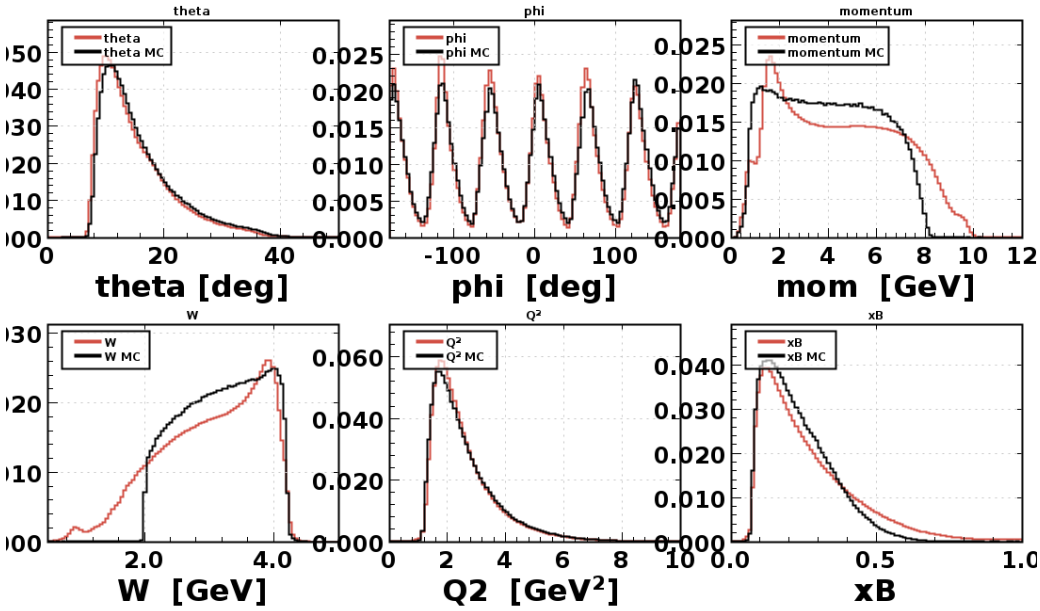

Fig. 39: Kinematics variables before cuts. Red lines are from RGA data and black lines are for the Monte-Carlo (MC) data.

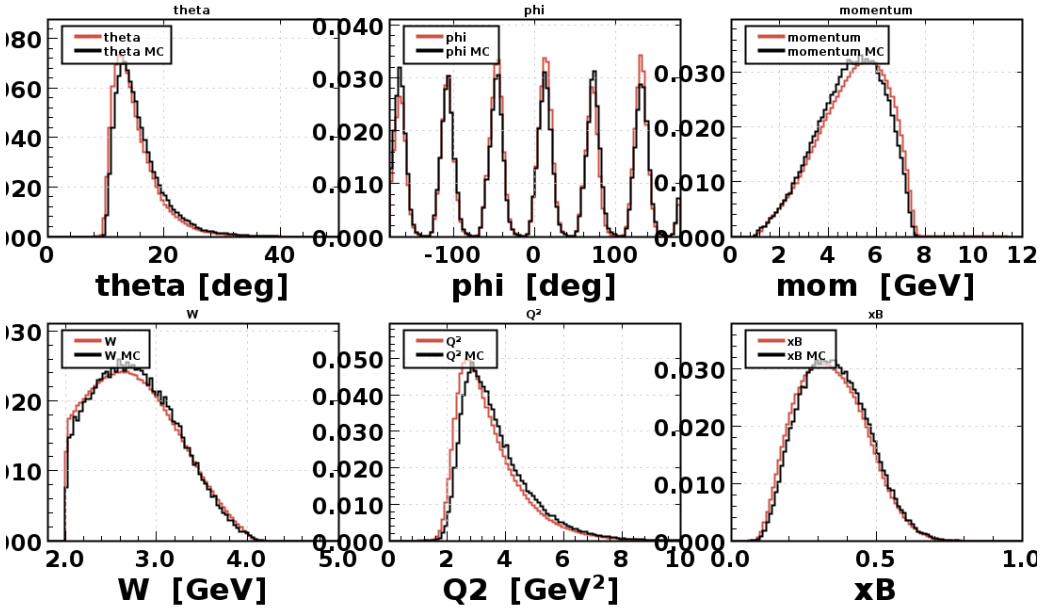

Fig. 40: Kinematics variables after all described cuts. Red lines are from RGA data and black lines are for the Monte-Carlo (MC) data.

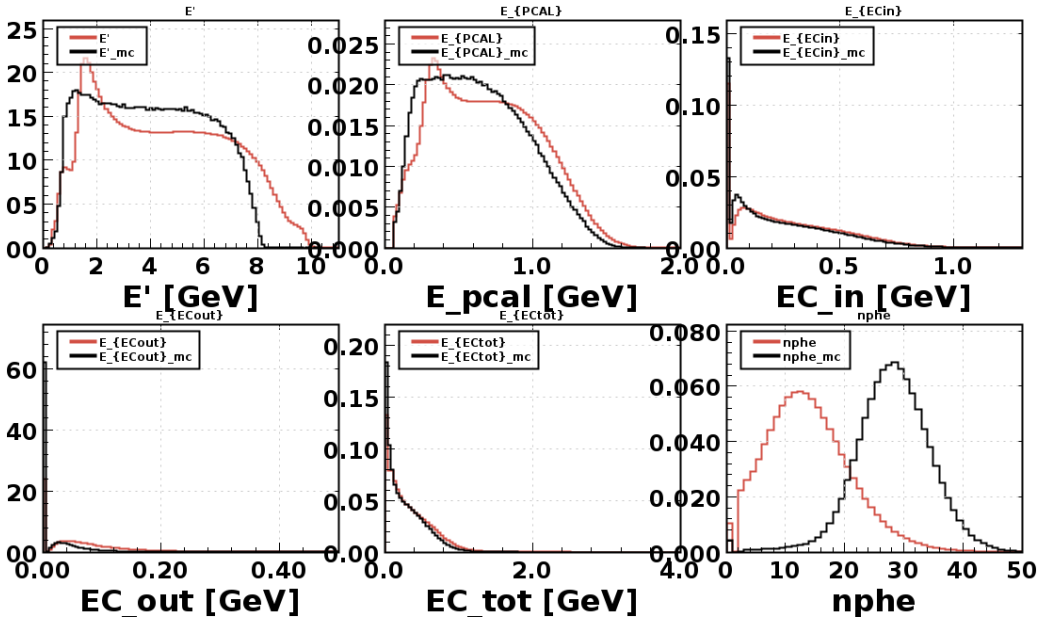

Fig. 41: E' and EC energies before cuts. Red lines are from RGA data and black lines are for the Monte-Carlo (MC) data.

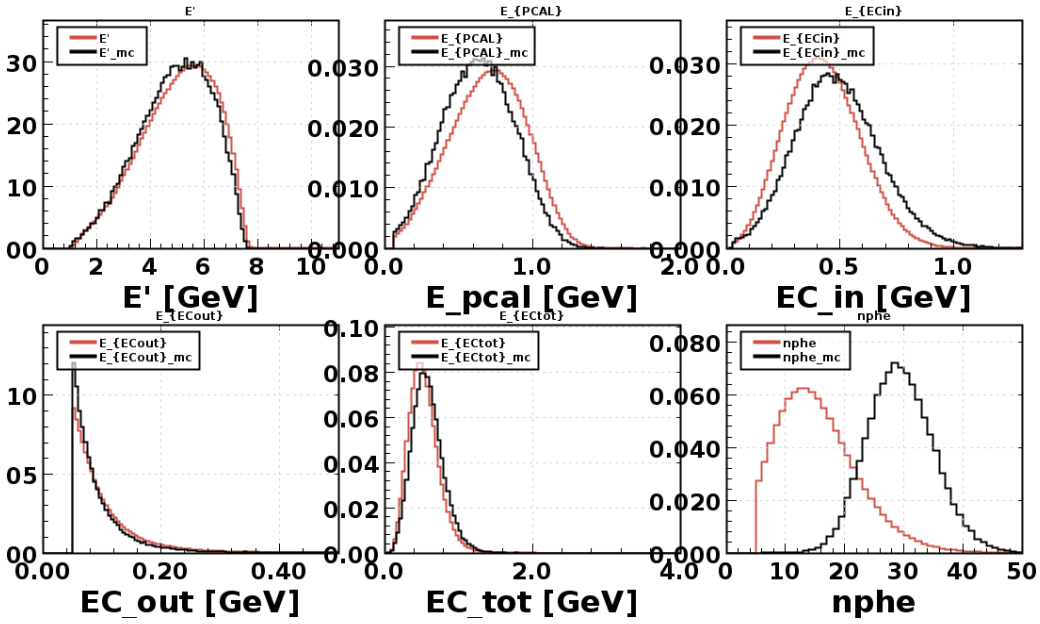

Fig. 42: E' and EC energies after all described cuts. Red lines are from RGA data and black lines are for the Monte-Carlo (MC) data.

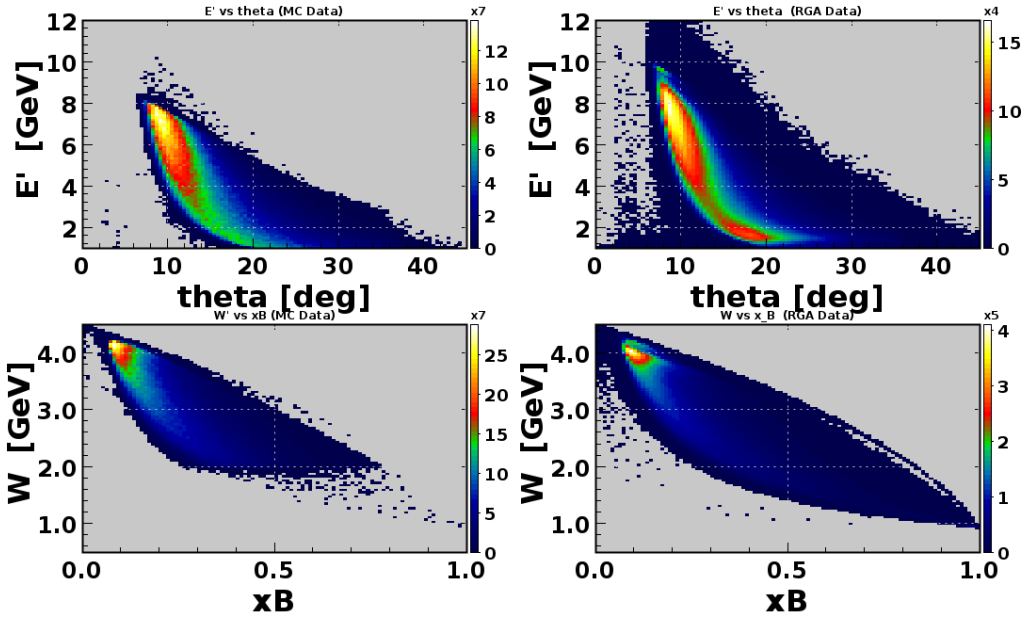

Fig. 43: E' vs. θ and W vs. x_B before cuts.

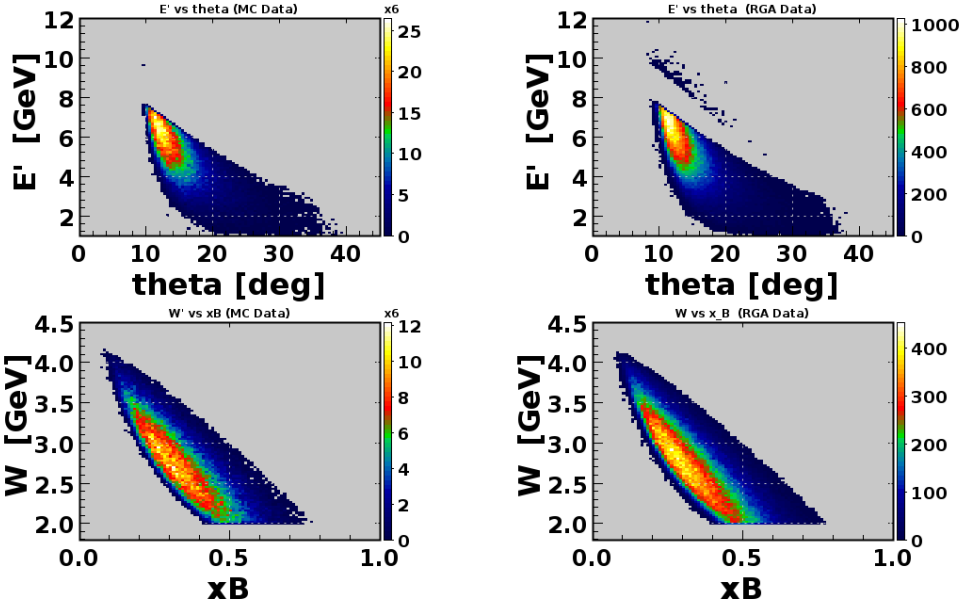

Fig. 44: E' vs. θ and W vs. x_B after all described cuts.

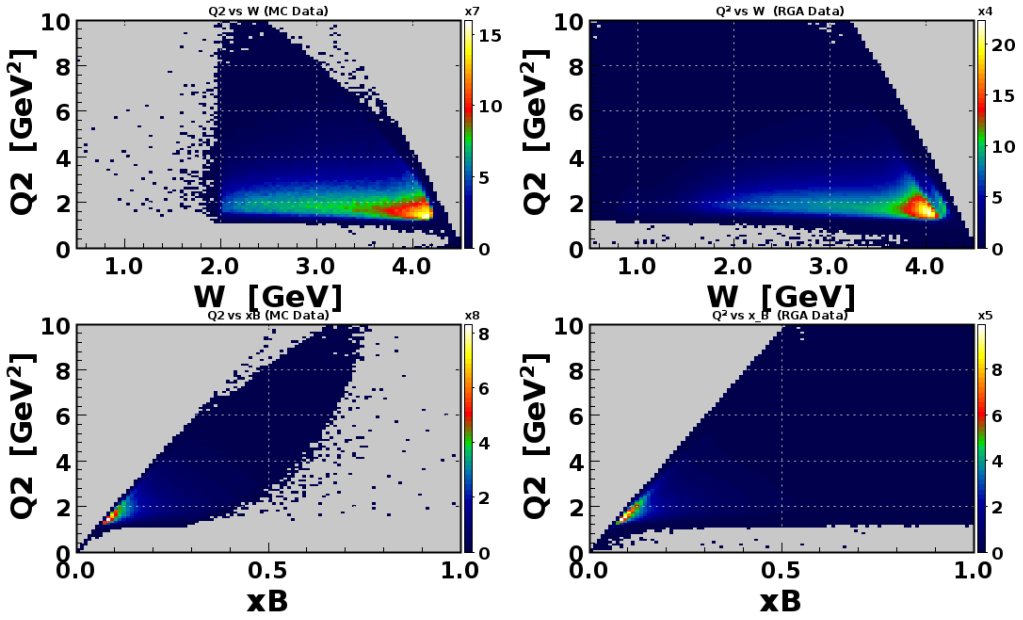

Fig. 45: Q^2 vs. W and Q^2 vs x_B before cuts.

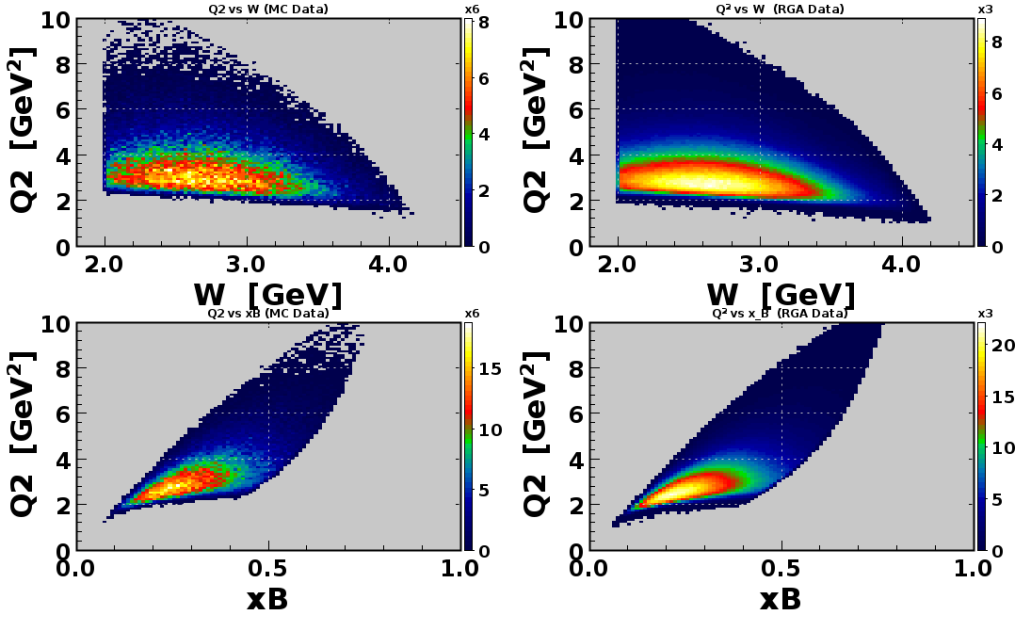

Fig. 46: Q^2 vs. W and Q^2 vs x_B after all described cuts.

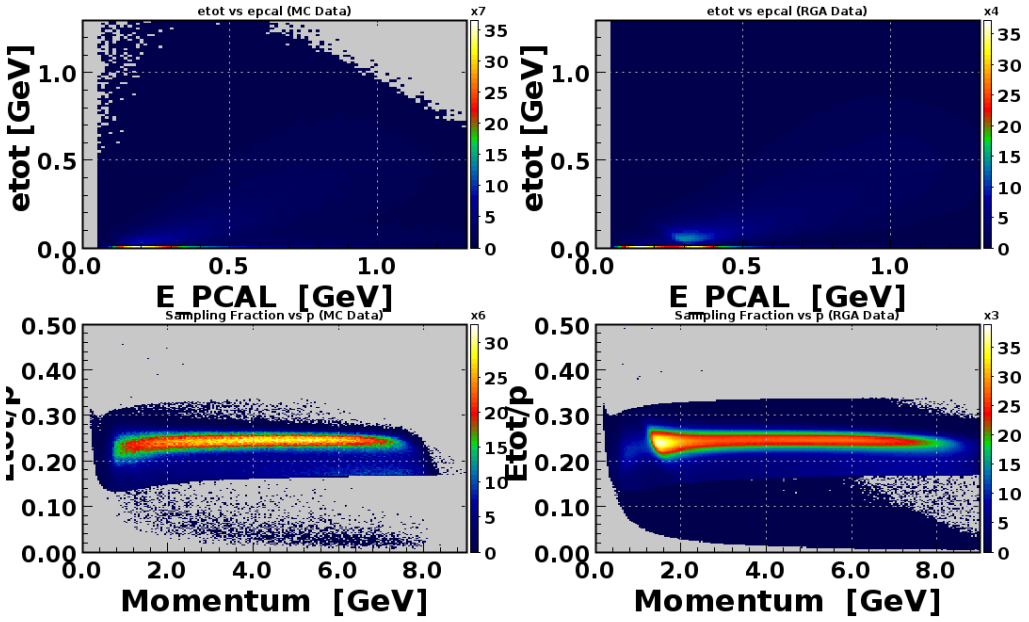

Fig. 47: $E_{EC_{tot}}$ vs E_{PCAL} and Sampling fraction before cuts.

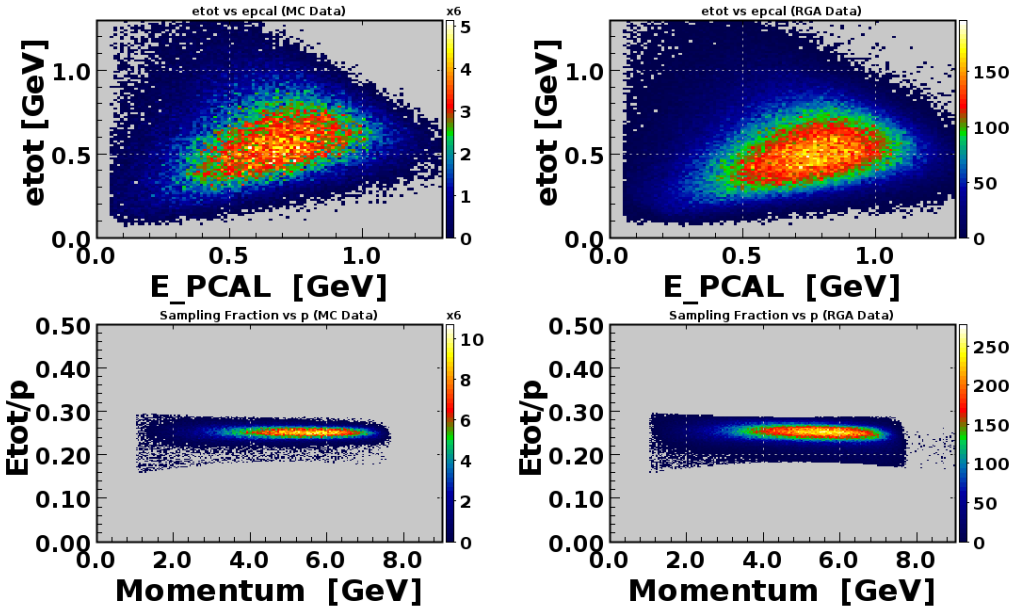

Fig. 48: $E_{EC_{tot}}$ vs E_{PCAL} and Sampling fraction after all described cuts.

5.5 BINNING AND ACCEPTANCE

The data, after kinematic and fiducial cuts, needs to be separated into kinematic bins in order to first obtain the acceptance of the bin and then to extract the cross section for each bin. The acceptance for the bin is defined as

$$A(x,y) = \frac{N_{\text{rec}}(x,y)}{N_{\text{gen}}(x,y)}, \tag{85}$$

where $A(x,y)$ is the acceptance of the bin, $N_{\text{gen}}(x,y)$ is the number of generated events in that bin, and $N_{\text{rec}}(x,y)$ is the number of reconstructed events in that bin. The binning occurs in x (*i.e.* the Bjorken-x scaling variable) and y, which is defined as $y = \nu/E$. This particular binning was done because of the cross section calculation that was done using model by Christy-Bosted, namely:

$$\frac{d^2\sigma}{dxdy} = \frac{4\pi\alpha_{\text{em}}S}{Q^4}\left[xy^2F_1(x,Q^2) + \left(1 - y - xy\frac{M^2}{S}\right)F_2(x,Q^2)\right], \tag{86}$$

where $S = 2ME$.

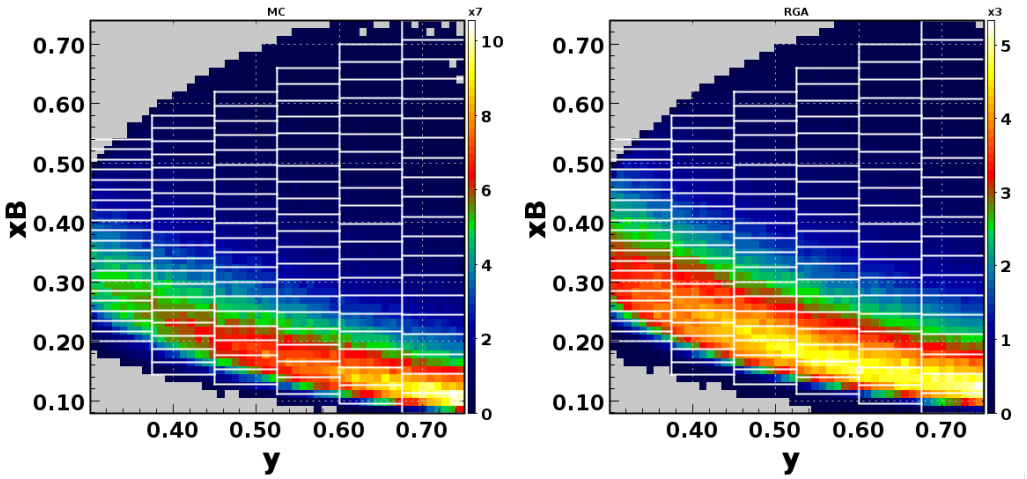

Fig. 49: Binning in the x, y space.

The binning in x and y was done in an attempt to gather equal statistics in each bin. Equal bins were chosen in y because, in the range $0.3 < y < 0.74$, the distribution was relatively flat. In x, however, the shape of the distribution required tuned binning throughout. Table IV outlines the bin size and range for x. Fig. 49 shows the binning in the x, y space.

Figs. 50-52 show the acceptance for each bin. For each bin in y, the value of which is located at the top of each plot, the acceptance is then plotted for each bin in x.

Bin	y Range	$x_{\min}$	$x_{\max}$	Δx
1	0.3, 0.375	0.2	0.54	0.17
2	0.375, 0.45	0.144	0.58	0.0218
3	0.45, 0.525	0.128	0.62	0.0246
4	0.525, 0.6	0.112	0.66	0.0274
5	0.6, 0.675	0.096	0.7	0.0302
6	0.675, 0.75	0.08	0.74	0.033

TABLE IV: Summary of binning in x and y.

Ideally the acceptance for all bins would be unity, so it is clear that the acceptance in areas of high and low y is lacking. In Fig. 50 we see that the acceptance is low at $x \to 0$ and in Fig. 52 acceptance drops as $x \to 1$.

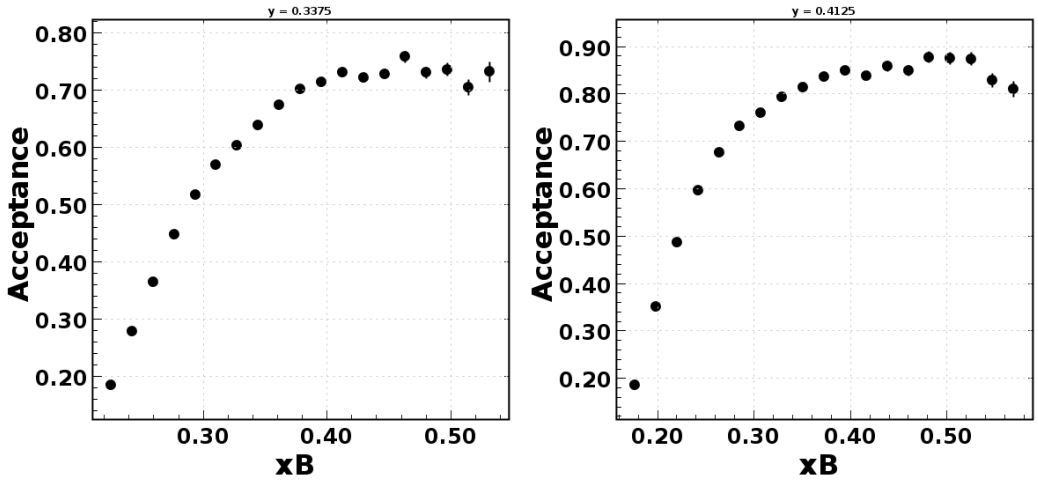

Fig. 50

5.6 FARADAY CUP AND INTEGRATED LUMINOSITY

As will become more evident in the next section, cross section calculations depend on the number of beam electrons accumulated during a particular run. Since that cross section is essentially the probability that a reaction occurs for a given process, it also depends on the number of target nuclei. The *luminosity* ($\mathscr{L}$) is a value incorporates both accumulated charge and number of target nuclei by expressing the number of beam particles per time multiplied by the number of target nuclei per unit

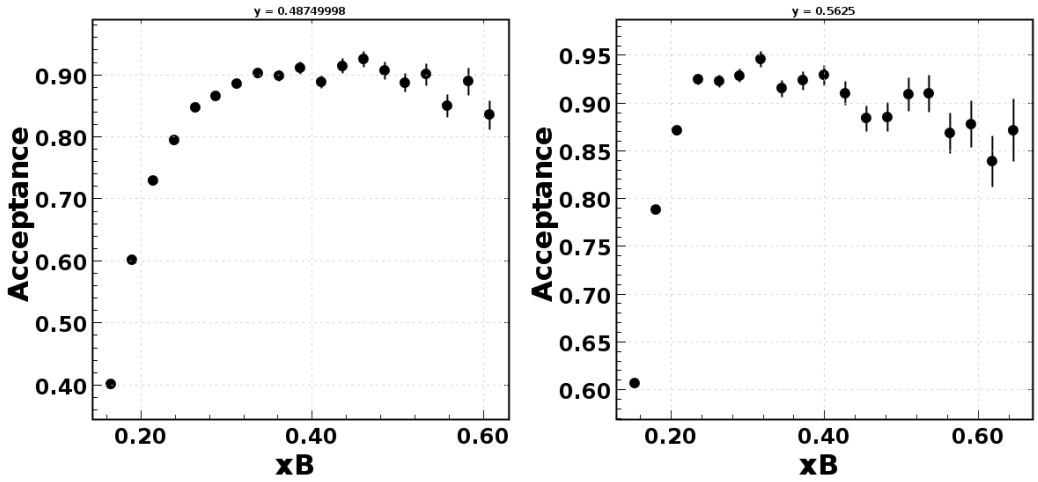

Fig. 51

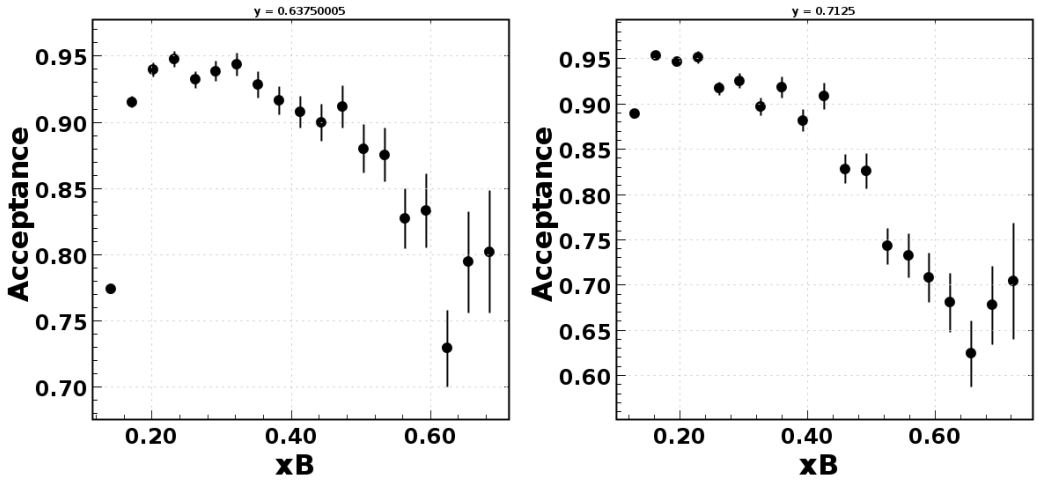

Fig. 52

area. By integrating that luminosity over time, we can recover the total number of beam electrons multiplied by the number of target nuclei per unit area

$$\mathscr{L}_{\text{int}} = \int \mathscr{L}\ dt = \frac{N_B \times N_{\text{target}}}{A}, \tag{87}$$

where N_B is the total number of incident electrons, N_{target} is the number of target nuclei, and A is the cross-sectional area of the target. This time-integrated luminosity ($\mathscr{L}_{\text{int}}$) depends on calculating N_{target}/A and knowing the total number of electrons incident on the target.

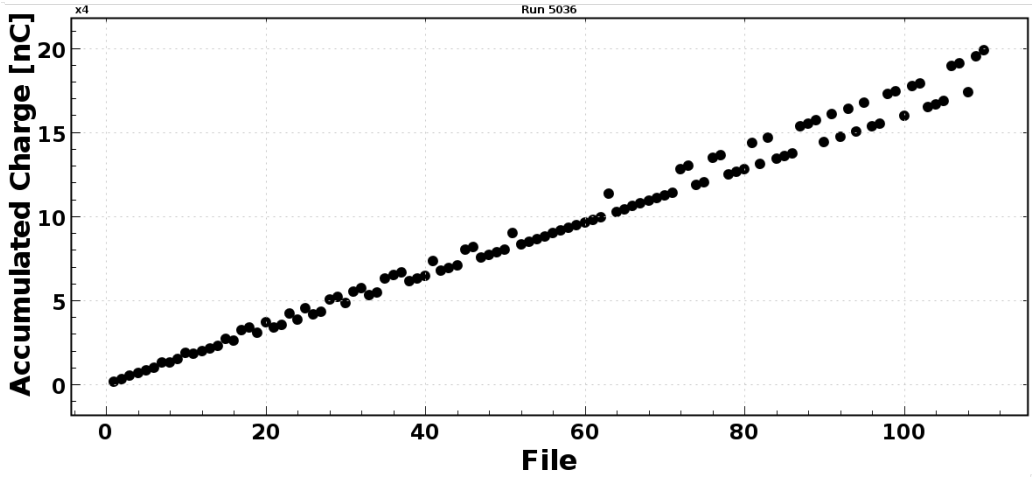

Fig. 53: Accumulated charge vs. file for Run 5036.

Calculating the number of target nuclei per area can be done utilizing the ideal gas law

$$PV = nRT, \tag{88}$$

where P is the target pressure, V is the volume of the target, n is the amount of target material that exists (in moles), R is the gas constant equal to 8.314472 J/K·mol, and T is the target temperature. We can rearrange the equation to get

$$N_{\text{target}} = 2nN_A = \frac{2PVN_A}{RT}, \tag{89}$$

where $N_A = 6.0221 \times 10^{23}$ mol^{-1} is Avogadro's number and the "2" comes from there being two hydrogen atoms in the liquid hydrogen target used for Run Group A (RGA). Finally, in order to find N_{target} in terms of the target density ρ, we use the relation

$$n = \frac{m}{M_m} = \frac{PV}{RT} \Rightarrow \rho = \frac{M_m P}{RT} \tag{90}$$

or

$$\frac{P}{T} = \frac{\rho R}{M_m}, \tag{91}$$

where M_m is the molar mass. That results in

$$N_{\text{target}} = \frac{2\rho V N_A}{M_m}. \tag{92}$$

This gives us a time-integrated luminosity

$$\mathscr{L}_{\text{int}} = \frac{2N_B N_A \ell \rho}{M_m}, \tag{93}$$

where ℓ is the length of the target.

The last variable to find is the total number of electrons N_B. We do this by accessing the charge accumulation in the Faraday Cup. The Faraday Cup (FC) is device located at the end of the beam line that catches charged particles, giving access to the total charge during a given period. The FC data is given in nano Coloumbs (nC) integrated over the entire run, where in every nC of charge there are 6.24150636309×10^9 electrons. Fig. 53 shows that accumulated charge as the number of files for the run 5036. That allows us to get the total number of incident electrons for each run, which is N_B in our integrated luminosity.

5.7 CROSS SECTION

The final step is to actually calculate the differential cross section for each x, y bin. That cross section for experimental data is given by

$$\frac{d^2\sigma}{dxdy} = \frac{N(x,y)}{\mathscr{L}_{\text{int}} A(x,y)\Delta x \Delta y}, \tag{94}$$

where $N(x,y)$ is the number of DIS events in the bin, $\mathscr{L}_{\text{int}}$ is the integrated luminosity, $A(x,y)$ is the acceptance of that bin, Δx is the size of the bin in x, and Δy is the size of the y bin. The number of selected inclusive deep inelastic scattering events $N(x,y)$ is calculated as the integral of the particular x, y bin.

Fig. 56 shows the calculated inclusive deep inelastic scattering differential cross sections (on the y-axis) for the values of y (listed as the title of the plots) and Bjorken-x on the x-axis. The open dots are the calculated cross sections from the RGA data. The green band is DIS cross sections calculated from Eq. 86 using Christy-Bosted fits of F_1 and F_2 with associated error and the same values of x and y for that bin. The error on the RGA calculated cross section is

$$\delta\frac{d^2\sigma(x,y)}{dxdy} = \frac{d^2\sigma/dxdy}{\sqrt{N(x,y)}}, \tag{95}$$

where $N(x,y)$ is the integral (or total number of entries) of the x, y bin.

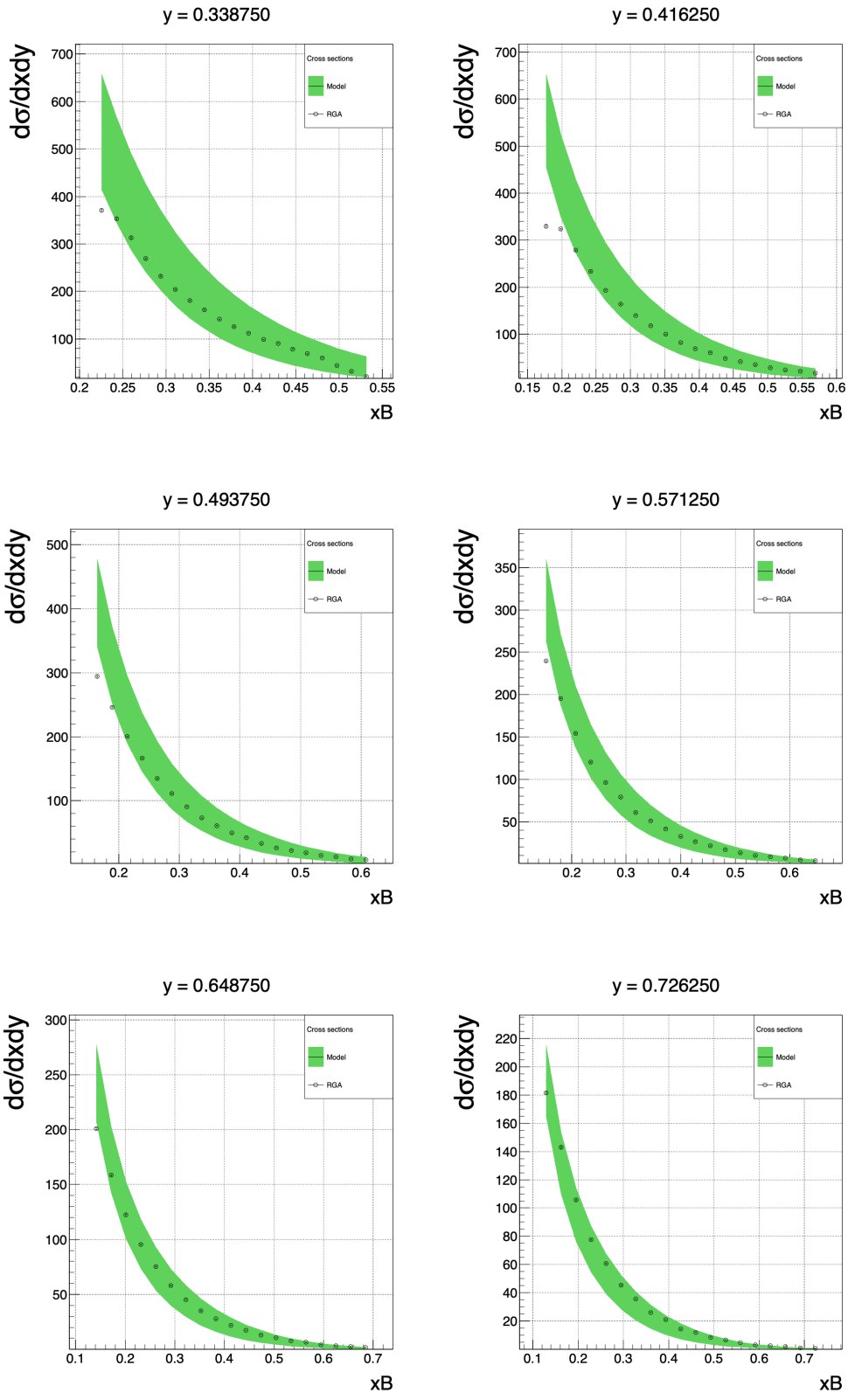

Fig. 56: Plots of inclusive IDS differential cross section vs x_B for various values of y.