UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Master Degree in Physics

Towards an Explainable Machine Learning Model of Proton Structure

Supervisor:
Prof. Stefano Forte
Co-supervisor:
Dr. Andrea Barontini

Student:
Martina Fusi
Matricola: 28494A

Year: 2023-2024

Abstract

Parton Distribution Functions (PDFs) are key tools in high-energy particle physics, essential for making predictions in Quantum Chromodynamics (QCD). The NNPDF collaboration has developed a method to determine PDFs using neural networks (NNs) and a Monte Carlo approach that generates an ensemble of equally probable replicas. The average of these replicas is known as the *central replica*.

While NNs are very successful at fitting the data, understanding the reasoning behind a particular trend in the PDFs is often difficult. This thesis aims to address this challenge by introducing explainable AI into the NNPDF framework.

In particular, we want to investigate the reasons behind the behaviour of the replicas of the gluon PDF. What happens is that replicas with best agreement with the data are distinctly separated from the ones with the worst agreement with the data: the first have a higher peak and a lower tail, while the second have a lower peak and higher tail. It was also found that the central replica is not the one with best agreement with the data. We want to understand why the replicas with the best agreement to the data prefer a higher peak and why this trend is not reflected by the central replica.

We analyzed datasets with the strongest correlations and anti-correlations between χ^2 and the peak height, finding that datasets favoring a higher peak contain more data points. By excluding these groups in separate fits, we confirmed that the peak's height and its correlation with total χ^2 are influenced by the excluded datasets, as expected.

In conclusion, the outlier replicas are driven by datasets that favor a higher peak. These replicas fit some data better and others worse, but their overall better agreement with the data is solely because the group of better-fitted datasets is larger.

Contents

Chapter 1

Introduction

We know since the Hofstadter experiments, in the 1950s, that the proton, unlike the electron, is not an elementary particle [2]. The concept of quarks arrived later, in 1964, postulated by Gell-Mann, more as a mathematical concept than as a physical prediction. By considering each hadron as made up of three quarks and each meson made up of two, it was possible to explain the zoo of new particles that were being discovered. A quark was a new kind of particle with fractional electric charge, baryon number, and spin–$\frac{1}{2}$.

In 1967, deep inelastic scattering (DIS) experiments at SLAC revealed that the proton-electron (p-e) scattering behaved like elastic scattering at high transferred momentum Q^2 [3]. This discovery, through the study of structure functions, led to the formulation of the parton model and consolidated the idea of quarks and, consequently, of parton distribution functions (PDFs).

PDFs are essential tools in high-energy particle physics, providing a framework for understanding the internal structure of protons and neutrons in terms of their constituent quarks and gluons. Accurate knowledge of PDFs is crucial for making predictions in quantum chromodynamics (QCD) processes and interpreting experimental results from particle colliders such as the Large Hadron Collider (LHC). A precise understanding of PDFs was essential for the discovery of the Higgs boson in 2012 and remains critical for uncovering new physics at the LHC. Indeed, uncertainties in PDFs remain one of the primary sources of error in numerous LHC processes.

In Figure 1.1, we show a notorious (though now outdated) depiction of the proton PDFs. In recent years, the NNPDF collaboration has developed and validated a method for the determination of PDFs which adopt neural networks (NNs). This method significantly improves the quality of the fit, removing the risk of biases. Furthermore, the NNPDF methodology uses a Monte Carlo sampling approach, where a large number of replicas are produced, and central values and uncertainties can be easily extract from the fit. This allows to have a reliable and solid value for the PDFs uncertainty, which can be out of reach for the methods which do not use NNs.
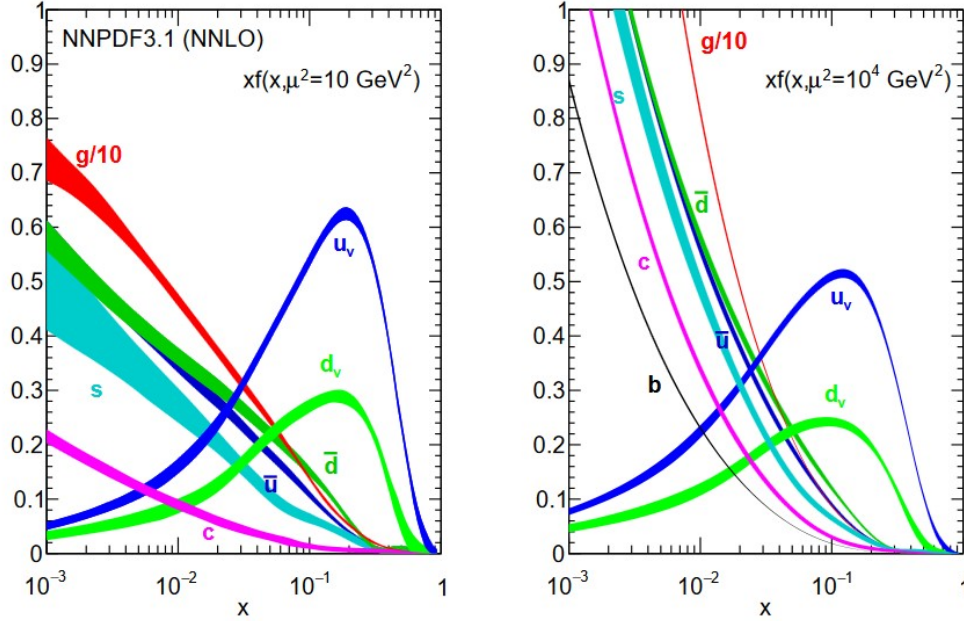
Figure 1.1: PDFs at next-to-next-to leading order (NNLO), obtained with NNPDF3.0 and evaluated at $\mu^2 = 10\,GeV^2$ (on the left) and $\mu^2 = 10^4\,GeV^2$ (on the right).

The flexibility provided by neural networks has made NNPDF a leading method for PDF determination, particularly because of its ability to accommodate large datasets and avoid bias from specific functional forms. However, with flexibility comes opacity: while NNs can fit the data extremely well, understanding the reasoning behind a particular fit or trend in the PDFs is often difficult. This lack of interpretability poses challenges when unexpected results arise.

This thesis aims to address this challenge by introducing explainable AI into the NNPDF framework. Explainable AI has emerged as a set of techniques aimed at making AI models more transparent and interpretable. The goal is to provide users with explanations for the decisions made by AI models.

In the context of PDF determination, explainability can help clarify why certain replicas behave as they do, revealing how different datasets contribute to the final outcome. In particular, we want to answer these questions for the gluon PDF.

The way in which the replicas of the gluon PDFs are distributed was firstly brought up in Ref.[32], where the authors noticed that the central replica of the fit does not coincide with the replicas with lowest χ^2. In particular, the replicas with the best χ^2 result to have a higher peak than the central replica. This, as later pointed out in Ref.[33], is not a worrying factor, since the best fit to the data is not necessarily the one with lowest χ^2. However, questions about why the replicas with lowest χ^2 prefer a higher gluon peak remain.

To answer these questions, we will investigate the correlation between the height of the

gluon peak and the χ^2 values on various datasets. This will allow to better understand which datasets drive specific trends in the replicas. This approach not only paves the way for more transparent PDF fits but also helps identify inconsistencies between different datasets, ultimately improving the reliability of future PDF determinations.

Structure of the Thesis

This thesis is structured as follows:

- **Chapter 2**: We review the theoretical background of PDF determination, with a focus on QCD and the experimental processes that provide the data used in the fits.

- **Chapter 3**: We discuss the methodology used in NNPDF4.0, focusing on how neural networks are used to parametrize the PDFs. This chapter also introduces the concept of k-fold cross-validation, used for hyperparameter optimization.

- **Chapter 4**: This chapter provides an overview of the datasets used in NNPDF4.0, detailing the different experimental sources, the different processes that contribute and the kinematic coverage they provide. We also list the complete set of datasets in Table 4.1, to which we will refer throughout the thesis.

- **Chapter 5**: We present the results of this thesis. We attempt to make the NNPDF fits more interpretable, highlighting which datasets drive specific trends in the PDF replicas. We perform fits excluding the highlighted datasets to better understand their impact.

- **Chapter 6**: We summarize the findings of the thesis, highlighting the most important results.

Chapter 2

Quantum Chromodynamics and Parton Distribution Functions

2.1 Quantum Cromodynamics (QCD)

QCD is a quantum field theory which describes the dynamics of the strong interaction in the Standard Model (SM). It is based on the local and non-abelian group SU(3), where the group gauge symmetry is associated with the conservation of the color current.

The strong interaction is mediated by gluons, which are massless, analogously to the photon, but, unlike the photon, they are not neutral. The number of gluons is equal to the dimension of the group adjoint representation, so we have a total number of 8 gluons. In fact, for a SU(N) groups, the dimension of the adjoint is $N^2 - 1$, so, given that we have three colors $N_c = 3$, the dimension is 8 [5].

We can represent each gluon with a basis vector of the adjoint representation; a commonly used basis are the Gell-Mann matrices λ^A, which are eight 3×3 traceless Hermitian matrices. Alternatively, one could also choose the matrices $t^A = \lambda^A/2$, which satisfy the algebra:

$$[t^A, t^B] = i\, f^{ABC} t^C, \tag{2.1}$$

where f^{ABC} are the structure constants.

The interaction among gluons can be found in the QCD classical Lagrangian:

$$\mathcal{L} = \sum_{i=1}^{n_f} \overline{\Psi}_i^a (i\gamma^\mu D_\mu - m_i)_{ab} \Psi_i^b - \frac{1}{4} F_{\mu\nu}^A F^{A\mu\nu}, \tag{2.2}$$

where the second term describe the interactions of massless spin-1 gluons, with $A = 1, ..., 8$ being the color index in the adjoint representation and μ, ν Lorentz indexes. $F_{\mu\nu}^A$ is the field

strength tensor, obtained as:

$$F_{\mu\nu}^A = \partial_\mu \mathcal{A}_\nu^A - \partial_\nu \mathcal{A}_\mu^A - g_s f^{ABC} \mathcal{A}_\mu^B \mathcal{A}_\nu^C, \tag{2.3}$$

with $\mathcal{A}_\mu^A$ being the gluon field and g_s the coupling, which is a free parameter of the theory. The last term, which is a consequence of QCD being a non-abelian theory, results in interactions between three and four gluons.

The first term of Eq.2.2 describes the evolution and interaction of spin-$\frac{1}{2}$ quark fields Ψ_i^a of mass m_i, where i runs over the n_f different flavors, while the indexes a, b refer to the color of the quark field. γ_μ are the usual gamma matrices, while D_μ is the covariant derivative, which, in order to rescue the local U(1) symmetry of the Lagrangian, is imposed to be:

$$D_\mu = \partial_\mu + ig_s(\mathcal{A}_\mu^A t^A), \tag{2.4}$$

where the term added to the usual derivative describes a quark-gluon interaction.

As said, the coupling constant g_s, often referred to as $\alpha_s = g_s^2/4\pi$, is a free parameter of the theory. However, its value is not fixed, but rather depends on, or runs with, the squared momentum transfer, Q^2:

$$\alpha_s(Q^2) = \frac{1}{\beta_0 \ \log(\frac{Q^2}{\Lambda^2})}, \tag{2.5}$$

where β_0 is a constant which depends on the symmetry group, μ^2 is the energy scale and Λ corresponds to the scale at which the coupling becomes infinite, also called Landau pole. In particular, if we are only considering the light quarks, $\beta_0 = (33 - 2\,n_f)/(12\pi)$.

The value of α_s increase at low energies/large distances ("soft physics", $\alpha_s \sim 1$) and decreases at high energy/small distances ("hard physics", $\alpha_s \sim 0.1$), as we can see in figure 2.1. This means that, as two quarks try to part, α_s increases, making the creation of a new quark-antiquark pair more energetically convenient than having free quarks.

QCD calculations are only possible at high energy, when α_s is sufficiently small and perturbation theory is applicable. This limit is also referred to as *asymptotic freedom*.

On the other hand, a comprehensive understanding of the low-energy behavior of the theory is still lacking.

2.2   The coupling constant α_s

The coupling function α_s sets the strength of the interactions involving quarks and gluons in QCD. Its determination is crucial, in fact, uncertainties in the value of $\alpha_s(Q^2)$ contribute significantly to the total theoretical uncertainty in the physics probed at the LHC.

In order to explain equation 2.5, we can consider a cross section $\sigma(s, \mu^2, \alpha_s)$, which depends on the invariant mass s, a renormalizing scale μ^2 and the coupling α_s. The constant α_s has
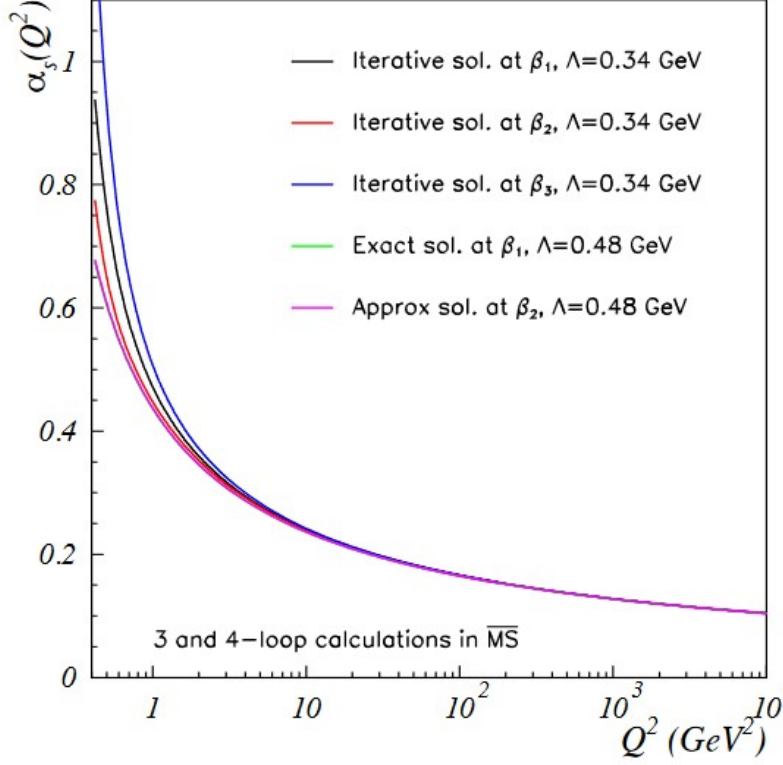
Figure 2.1: The perturbative coupling $\alpha_s(Q^2)$ computed at different orders in β_i and for two different values of Λ. The coefficients β_i are computed in the $\overline{MS}$. This figure is from ref.[1].

to be redefined to absorb the dependence on the renormalization scale μ^2, so we will now call it $\alpha_s(\mu^2)$. The cross section is then $\sigma(s, \mu^2, \alpha_s(\mu^2))$.

It is important to stress that physical predictions, such as the total cross section σ, can not depend on the renormalization scale μ^2. This conduct us to the *renormalization group equation*:

$$\mu^2 \frac{\mathrm{d}}{\mathrm{d}\mu^2} \sigma(\frac{s}{\mu^2}, \alpha_s(\mu^2)) = 0 = -s \frac{\partial}{\partial s} \sigma\left(\frac{s}{\mu^2}, \alpha_s(\mu^2)\right) + \mu^2 \frac{\partial \alpha_s(\mu^2)}{\partial \mu^2} \frac{\partial}{\partial \alpha_s(\mu^2)} \sigma\left(\frac{s}{\mu^2}, \alpha_s(\mu^2)\right) \quad (2.6)$$

looking at the right side, this gives us the important operator relation:

$$s \frac{\mathrm{d}}{\mathrm{d}s} = \mu^2 \frac{\mathrm{d}}{\mathrm{d}\mu^2} \quad (2.7)$$

We now define the *beta function* $\beta(\alpha_s(\mu^2)) \equiv \mu^2 \frac{\mathrm{d}\alpha_s(\mu^2)}{\mathrm{d}\mu^2}$, which can be written as

$$\beta(\alpha_s(\mu^2)) = -b_0 \alpha_s(\mu^2)^n + b_1 \alpha_s(\mu^2)^{n+1} + \dots \quad (2.8)$$

9

From equation 2.7 we can rewrite $\beta(\alpha_R) = \frac{d}{d\ln s}\alpha_R(s)$, where α_R is the running coupling constant with boundary condition $\alpha_R(\mu^2) = \alpha_s(\mu^2)$. We can then solve the differential equation:

$$\int_{\mu^2}^{s} d\ln s = \int_{\alpha_R(\mu^2)}^{\alpha_R(s)} \frac{d\alpha_R}{\beta(\alpha_R)} \tag{2.9}$$

taking $\beta(\alpha_R) = -b_0\alpha_R^2$, the solution is:

$$\ln\frac{s}{\mu^2} = \frac{1}{b_0\alpha(s)} - \frac{1}{\alpha(\mu^2)} \tag{2.10}$$

which leads us to:

$$\alpha_R(s) = \frac{\alpha_R(\mu^2)}{1 + b_0\alpha_R(\mu^2)\ln\frac{s}{\mu^2}} \tag{2.11}$$

This beta function is a property of the theory, with b_0 independent on the chosen renormalization scheme, while the other coefficients depend, in general, on the scheme. In QCD $b_0 > 0$, hence we have the asymptotic freedom at large s and the growth of g at small s. We can go as far as defining a value Λ, at which the running constant goes to infinity. This is defined as:

$$1 + b_0 g(\mu^2)\ln\frac{s}{\mu^2} = 0 \tag{2.12}$$

With this definition, we obtain equation 2.5:

$$g_R(s) = \frac{1}{b_0 \ln\frac{s}{\Lambda^2}} \tag{2.13}$$

2.3 Deep Inelastic Scattering

Deep inelastic scattering (DIS) is one of the cornerstone in PDFs determination. As we saw in the introduction, it was also the first indication of hadron internal structure.

DIS consist in lepton-hadron scattering, where the lepton is highly energetic and can be both charged or neutral. An example of DIS is shown in Figure 2.2. The interaction takes place by means of the interchange of a boson. If the boson is a photon $\gamma*$ or a Z boson, the process is called neutral DIS. If the boson is a $W^\pm$, the process is called charged DIS.

It is convenient to define the following invariant quantities:

$$Q^2 \equiv -q^2, \quad x_{bj} \equiv \frac{Q^2}{2P\cdot q}, \quad y \equiv \frac{q\cdot P}{k\cdot P} \tag{2.14}$$

where we labelled P the momentum of the incoming hadron, q is the momentum of the exchanged gauge boson and k the momentum of the incoming lepton. Q^2 is the virtuality of the gauge boson and y is the inelasticity. The invariant x_{bj} is known as Bjorken variable and it is equal to the fraction of the hardon momentum taken by the parton in the parton model.
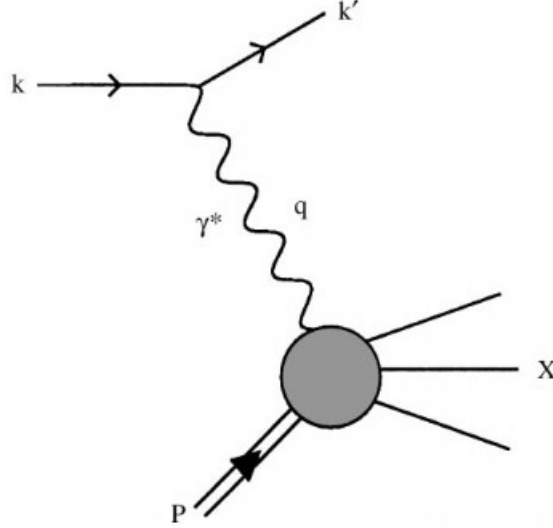
Figure 2.2: Representation of a neutral DIS process with the exchange of a virtual photon γ^*.

In QCD, this holds only at leading order (LO).

We call $p = xP$ the momentum of the incoming parton, where P is the total hadron momentum. We want to prove the equivalence between x_{bj} and the fraction of hadron momentum x in the parton model.

We will start computing the cross section at tree level of the interaction between a proton and a photon. But first, we recall that a cross section is the integration of the differential cross section $d\sigma$ over the solid angle $d\Omega$:

$$\sigma = \int d\Omega \frac{d\sigma}{d\Omega}, \tag{2.15}$$

where the differential cross section is obtained as:

$$d\sigma = \frac{1}{flux} |\overline{\mathcal{M}}|^2 d\Phi. \tag{2.16}$$

The matrix element $\mathcal{M}$ is calculated through Feynman diagrams, while the flux factor and the phase space invariant $d\Phi$, when we have a $2 \to 1$ scattering with in-coming momentum p_A and p_B and out-coming momentum p_C, are:

$$flux = 4\sqrt{(p_A \cdot p_B)^2 - m_A^2 m_B^2}, \tag{2.17}$$

$$d\Phi = (2\pi)^4 \delta^4(p_A + p_B - p_C) \frac{d^3 p_C}{(2\pi)^3 2E_C}. \tag{2.18}$$

When computing the tree level cross section of a proton-photon process, we have to take

into account all the possible sub-processes occurring between the photon and each of the quarks, so the cross section will be:

$$\sigma_0 = \int d\Omega \int_0^1 dx \sum_i Q_i^2 \, f_i(x) \, \frac{d}{d\Omega} \hat{\sigma}_0(\gamma^* q_i \to q_i', \, x), \tag{2.19}$$

where Q_i is the charge of the quark of flavour i, $f_i(x)$ gives the probability of having a quarks of flavour i carrying a momentum xP, where P is the proton total momentum and x is the fraction of hadron momentum. Lastly, $\hat{\sigma}_0(\gamma^* q_i \to q_i', \, x)$ is the cross section of the subprocess involving the quark of flavour i. This step is also known as *factorization*.

This cross section is obtained by integrating the differential cross section of Equation 2.20. Using the properties of the Dirac delta function, it can be shown (though we omit the detailed steps here) that the phase space invariant can be expressed as:

$$(2\pi)^4 \int \frac{d^3 p'}{(2\pi)^3 2p_0'} \, \delta^4(p' - xP - q) = 2\pi \delta(k_0'^2), \tag{2.20}$$

where p' is the four-momentum of the outgoing quark and q is the momentum of the photon.

Now, knowing that:

$$p'^2 = (xP + q)^2 = 2xP \cdot q + q^2 = 2xP \cdot q - Q^2, \tag{2.21}$$

and therefore, using the Dirac δ properties we can write the differential cross section as:

$$\frac{d\hat{\sigma}_0}{d\Omega} = \frac{2\pi}{flux} \, \overline{\mathcal{M}_0^2} \, \frac{1}{2P \cdot q} \delta(x - x_{bj}), \tag{2.22}$$

where $x_{bj} = Q^2/2P \cdot q$, as defined above. This shows the equivalence between x and x_{bj} in the parton model.

Substituting this in Equation 2.19, we obtain (with an integrating over x):

$$\sigma_0 = \int d\Omega \, \frac{2\pi}{flux} \, \frac{\overline{\mathcal{M}_0^2}}{Q^2} \sum_i x_{bj} \, Q_i^2 \, f_i(x) = \int d\Omega \, \frac{2\pi}{flux} \, \frac{\overline{\mathcal{M}_0^2}}{Q^2} \, F(x_{bj}), \tag{2.23}$$

where we have introduced the structure function:

$$F(x) = \sum_i x_{bj} Q_i^2 f_i(x). \tag{2.24}$$

The final outcome is that, measuring the differential cross section of a proton-electron scattering, we obtain important information about the structure function $F(x)$, and thus the quark PDFs $f_i(x)$.

2.3.1 Parton Evolution

We now want to take into consideration the first order correction of the cross section that we have just found. This means considering the emission of a real gluon. The corresponding Feynman diagrams are shown in Figure 2.3.
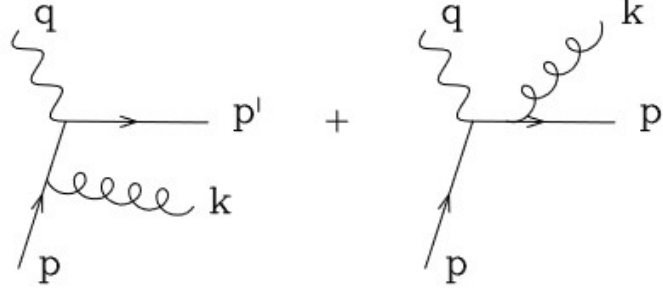

Figure 2.3: QCD corrections to tree-level DIS.

These diagrams carry a divergence when the gluon is emitted parallel to the quark. In the right diagram of Figure 2.3, this divergence turns out to be harmless, since it can be cancelled by a similar divergence caused by the final-state quark self-energy corrections. So, as long as we are inclusive on the final state, this divergence is not problematic.

The situation is different in the case of a gluon emitted by the incoming quark. This divergence can not be cancelled by other corrections and it has to be treated in a different way. This problem requires that the PDFs $f_i(x)$ acquire a dependence on the scale. In this way, the divergence can be reabsorbed in the PDFs, and physical observables, such as the cross section σ, can be calculated with finite results.

This requires the introduction of a scale μ^2. Since the value of μ^2 is arbitrary, the total cross section $\sigma(x, Q^2)$, should not depend on it. So, we impose:

$$\frac{d\sigma}{d\log\mu^2} = 0. \tag{2.25}$$

Since the scale dependency has to cancel between $f_i(x)$ and σ, and σ is calculated with the PDFs $f_i(x)$, Equation 2.25 leads us to the DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) equation:

$$\mu^2 \frac{df(x, \mu^2)}{d\mu^2} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \, P_{qq}(z) \, f_i\left(\frac{x_{bj}}{z}, \mu^2\right), \tag{2.26}$$

where P_{qq} is the so-called Aterelli-Parisi splitting function for the $q \to q$ transition, while z is the momentum fraction of the initial quark taken away by the quark after the gluon emission.

This result depends on the parton, for the splitting of a gluon into a $q\bar{q}$ pair, we would

obtain a different solution, with an extra term:

$$\frac{df_q(x,t)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \left[P_{qq}(z) f_i(\frac{x_{bj}}{z}, t) + P_{qg}(z) f_g\left(\frac{x_{bj}}{z}, t\right) \right] \tag{2.27}$$

where $t = \log(\frac{Q^2}{\mu^2})$ and $P_{qg} = \frac{1}{2}[z^2 + (1-z)^2]$. The same thing holds for the evolution of the gluon, this time we also have to consider the decay of the gluon in a quark-antiquark pair and the production of a second gluon:

$$\frac{df_g(x,t)}{dt} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \left[P_{gq}(z) \sum_{i=q,\bar{q}} f_i\left(\frac{x_{bj}}{z}, t\right) + P_{gg}(z) f_g\left(\frac{x_{bj}}{z}, t\right) \right]. \tag{2.28}$$

These equations provide a description of the partons evolution for processes at next-to-leading order (NLO). These equations can also be summarize in:

$$Q^2 \frac{\partial}{\partial Q^2} f_i(x, Q^2) = \sum_j P_{ij}(x, \alpha_s(Q^2)) \otimes f_j(x, Q^2), \tag{2.29}$$

where $\otimes$ denotes the convolution:

$$(f \otimes g)(x) \equiv \int_1^x \frac{dy}{y} f(y) g\left(\frac{x}{y}\right). \tag{2.30}$$

2.3.2 Solution of the evolution equations

We introduce the Mellin transform:

$$g^{(n)} = \int_0^1 \frac{dx}{x} x^n g(x), \tag{2.31}$$

with this, we can rewrite the evolution equations as linear differential equations [5]:

$$\frac{df_i^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [\gamma_{qq}^{(n)} f_i^{(n)} + \gamma_{qg}^{(n)} f_g^{(n)}], \tag{2.32}$$

$$\frac{df_g^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [\gamma_{gg}^{(n)} f_i^{(n)} + \gamma_{gq}^{(n)} f_g^{(n)}], \tag{2.33}$$

where γ_{ij} is the Mellin trasform of the splitting function P_{ij}, which is computed order by order in perturbation theory:

$$P_{ij} = \alpha_s P_{ij}^{(0)} + \alpha_s^2 P_{ij}^{(1)} + \dots \tag{2.34}$$

The splitting functions at LO in x-space can be found in Table 2.1.

There are $n_f + 1$ coupled equations, however, we can reorganize these equations into the
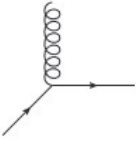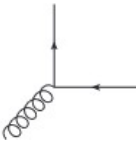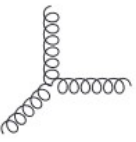
Diagram	Splitting Function
	$$P_{qg}^{(0)}(x) = C_F \left[\frac{1+x^2}{(1-x)_+} + \frac{3}{2}\delta(1-x) \right]$$
	$$P_{gq}^{(0)}(x) = C_F \left[\frac{1+(1-x)^2}{x} \right]$$
	$$P_{qg}^{(0)}(x) = T_F \left[x^2 + (1-x)^2 \right]$$
	$$P_{gg}^{(0)}(x) = 2C_A \left[\frac{1-x}{x} + x(1-x) + \frac{x}{1-x} \right] + \delta(1-x)\frac{11C_A - 4T_F n_f}{6}$$

Table 2.1: Splitting functions at LO [39]. C_F, T_F and C_A are numerical quantities associated with the SU(3) groups and have value 4/3, 1/2 and 3, respectively.

following combinations (here we define them below the charm threshold):

$$\Sigma(x, Q^2) \equiv \sum_{i=1}^{n_f} (q_i + \bar{q}_i)(x, Q^2),$$

$$T_3(x, Q^2) \equiv (u + \bar{u} - d - \bar{d})(x, Q^2),$$

$$T_8(x, Q^2) \equiv (u + \bar{u} + d + \bar{d} - 2(s + \bar{s}))(x, Q^2),$$

$$V(x, Q^2) \equiv \sum_{i=1}^{n_f} (q_i - \bar{q}_i)(x, Q^2),$$

$$V_3(x, Q^2) \equiv (u - \bar{u} - d + \bar{d})(x, Q^2),$$

$$V_8(x, Q^2) \equiv (u - \bar{u} + d - \bar{d} - 2(s - \bar{s}))(x, Q^2).$$

(2.35)

We now look at valence $(V(x, t))$, singlet $(\Sigma(x, t))$ and gluon density $(f_g(x, t))$. The evolution equations with this redefinition become:

$$\frac{dV^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [\gamma_{qq}^{(n)} V^{(n)}],$$

(2.36)

$$\frac{d\Sigma^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [\gamma_{qq}^{(n)} \Sigma^{(n)} + 2 n_f \gamma_{qg}^{(n)} f_g^{(n)}],$$

(2.37)

$$\frac{df_g^{(n)}}{dt} = \frac{\alpha_s}{2\pi} [\gamma_{gg}^{(n)} f_i^{(n)} + \gamma_{gq}^{(n)} f_g^{(n)}].$$

(2.38)

The importance of this basis in that it maximally decouples the differential equations. In fact, the valence V (among with V_3, V_8, T_3 and T_8) is decoupled from Σ and f_g. Only Σ and f_g remain coupled.

With this basis, we can easily solve the differential equation for V at LO. We take into account the running of α_s (Eq. 2.5):

$$\frac{dV^{(0)}}{dt} = \frac{1}{2\pi b_0 \log(\frac{Q^2}{\Lambda^2})} [\gamma_{qq}^{(0)} V^{(0)}] = \frac{1}{2\pi b_0 [t + \log(\frac{\mu^2}{\Lambda^2})]} [\gamma_{qq}^{(0)} V^{(0)}]$$

(2.39)

which we can integrate using as boundary conditions $V(\mu^2)^{(0)}$ (with $\mu^2 >> Q^2$):

$$\int_{V^{(0)}(\mu^2)}^{V^{(0)}(Q^2)} \frac{1}{V^{(0)}} dV^{(0)} = \int_0^{\log \frac{Q^2}{\mu^2}} \frac{dt}{2\pi b_0 (t + \log \frac{\mu^2}{\Lambda^2})} \gamma_{qq}^{(0)},$$

(2.40)

solving which we obtain:

$$V^{(0)}(Q^2) = V^{(0)}(\mu^2) \left(\frac{\log \frac{Q^2}{\Lambda^2}}{\log \frac{\mu^2}{\Lambda^2}} \right)^{\frac{\gamma_{qq}^{(0)}}{2\pi b_0}} = V^{(n)}(\mu^2) \left(\frac{\alpha_s(\mu^2)}{\alpha_s(Q^2)} \right)^{\frac{\gamma_{qq}^{0)}}{2\pi b_0}}$$

(2.41)

16

We conclude that at LO the valence quarks only evolve due to the emission of gluons. This also tells us the dependence of the valence on Q^2: since $P_{qq} < 0$, the valence quarks will have less and less energy (on average) when Q^2 increases. The left panel of Fig. 2.4 shows this effect. We can notice that the peak of the distribution is in correspondence of smaller values of x as Q^2 increases.
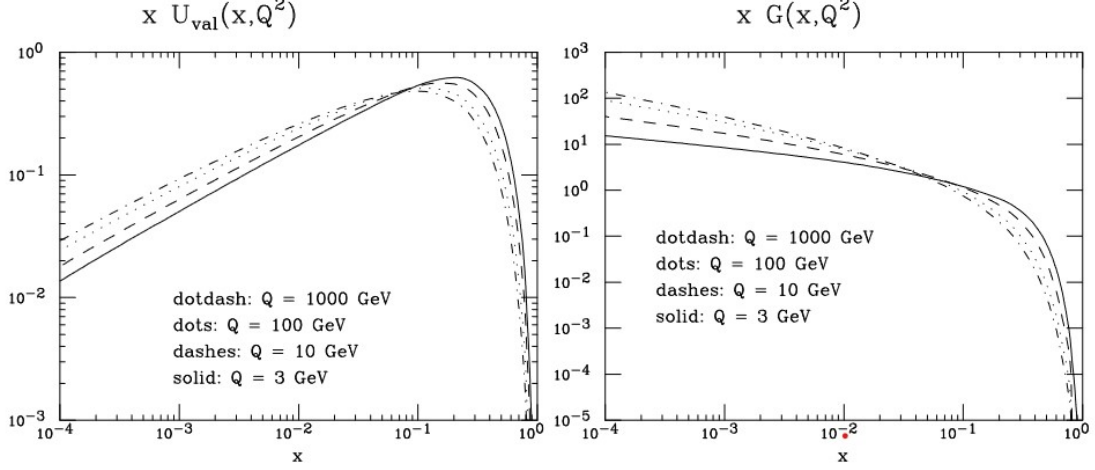

Figure 2.4: On the left: up quark valence momentum density at different values of Q^2; on the right: gluon momentum density at different scales Q^2.

To study the behaviour of Σ and f_g, it is convenient to notice that the sum of the momentum of the constituents of the proton (sea quarks, valence quarks and gluons) should always give the total momentum of the hadron, regardless of Q^2. This implies:

$$\int_0^1 dx\, x \left[\sum_{i,\bar{i}} f_i(x) + f_g(x) \right] = \Sigma^{(2)} + f_g^{(2)} = 1 \tag{2.42}$$

and it is called the *momentum sum* rule. Using this property together with the fact that $\frac{d\Sigma^{(2)}}{dt} = 0$ in the asymptotic limit, we obtain the system:

$$\begin{cases} \gamma_{qq}^{(2)} \Sigma^{(2)} + 2n_f \gamma_{qg}^{(2)} f_g^{(2)} = 0 \\ \Sigma^{(2)} + f_g^{(2)} = 1 \end{cases} \tag{2.43}$$

which we can solve with the use of $P_{qq} = C_F \frac{1+z^2}{1-z}$ and $P_{qg} = \frac{1}{2}[z^2 + (1-z)^2]$:

$$\Sigma^{(2)} = \frac{1}{1 + \frac{4C_F}{n_f}} \sim \frac{1}{2}, \tag{2.44}$$

$$f_g^{(2)} = \frac{4C_F}{4C_F + n_f} \sim \frac{1}{2}, \tag{2.45}$$

17

where $C_F = 4/3$ and the number of flavors is $n_f = 5$ (we don't consider the top quark). The outcome of this calculation is that gluons carry, asymptotically, around 50% of the proton momentum.

On the right panel of Figure 2.4, we have the gluon momentum density. We can notice that, while the valence momentum density vanishes for $x \to 0$, f_g increases at low values of x. This comes from the fact that both P_{gg} and P_{gq} have a $\frac{1}{x}$ dependence.
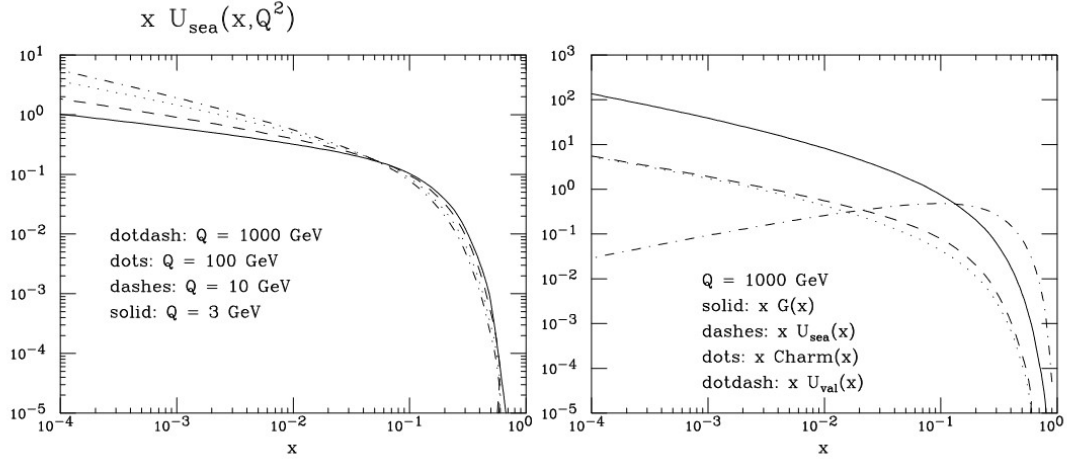

Figure 2.5: On the left: sea up quark momentum density at different values of Q^2; on the right: momentum density distribution of different partons at $Q^2 = 1000 \, GeV$.

In Figure 2.5, the momentum density distribution of the sea up-quark is shown on the left panel, along with momentum density distribution of different partons in the proton, on the right panel. All the parton distribution are scale dependent (they depend on Q^2).

From the summary plot (right panel of fig.2.5), we can deduce that (at $Q^2 = 1000 \, GeV$) valence quarks and gluons carry the majority of the proton momentum, while sea quarks carry the least momentum, which is often negligible.

2.3.3   Constraints on Proton PDFs

Before moving on to the next section, it is important to sum up the most important results we obtained so far. In this section we want to highlight the theoretical constraints on the proton's PDFs, which will be very useful later.

We are not able to solve QCD and compute PDFs at present, so they need to be deduced from data. Despite this, there are some theoretical constraints that must be satisfied.

Since the PDFs must yield the quantum numbers that characterize the proton, the integral of the valence down-quark $d_v(x)$ should add up to 1, whereas the integral of the valence up-quarks should be 2. These, simplifying the notation by referring to $f_i(x)$ ($i = u, d, \overline{u}, \overline{d}...$) as

simply $i(x)$, are called *valence sum* rules:

$$\int_0^1 dx \left(d(x, Q^2) - \bar{d}(x, Q^2) \right) = \int_0^1 dx \, d_v(x) = 1, \qquad (2.46)$$

$$\int_0^1 dx \left(u(x, Q^2) - \bar{u}(x, Q^2) \right) = \int_0^1 dx \, u_v(x) = 2. \qquad (2.47)$$

The three valence quarks carry proton charge $(+1)$ and baryon number $(+1)$. We recall that the proton is also made of light quark pairs $q\bar{q}$ and gluons.

The second theoretical constraint is the *momentum sum rule*, previously derived in eq.2.42:

$$\sum_{i=q,\bar{q},g} \int_0^1 dx \, x f_i(x, Q^2) = 1.$$

This provides the condition that the sum of the momentum of the constituents should be equal to the momentum of the proton.

A further requirement on the PDFs is that they should vanish as $x \to 1$, since $f(x) = 0$ for $x > 1$ is garanteed by energy momentum conservation:

$$f_i(x = 1, Q) = 0, \qquad (2.48)$$

At the same time, the valence sum rules require the corresponding distributions to be integrable on the entire range in x.

2.4   QCD in Hadronic Collisions

In hadronic collisions, such as proton-antiproton ($p\bar{p}$), we may wonder whether factorization still holds. It can be proven that, for sufficiently inclusive quantities, this is the case. So, for hadronic collision processes, the differential cross section is given by:

$$d\sigma(p\bar{p} \to A + X, \, s) = \int dx_1 \, dx_2 \sum_{i,j} f_i(x_1, Q) \, f_j(x_2, Q) \, d\hat{\sigma}(ij \to A + X, \, x_1 x_2 s), \qquad (2.49)$$

where A is the derived final state (e.g. a Higgs or Z boson), X are the other residual particles and $d\hat{\sigma}(ij \to A)$ is the differential cross section of the process involving the partons i and j, respectively from the proton and antiproton. The invariant mass of the sub-process is $\hat{s} = x_1 x_2 s$, where s is the invariant mass of the hadronic process. Lastly, the value of Q, at which f_i and f_j are evaluated at, is at a scale typical of the hard process they take part in.
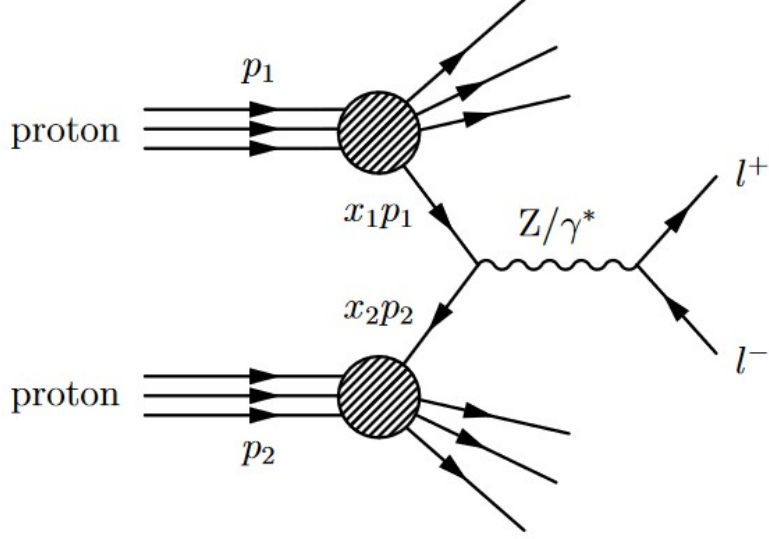
Figure 2.6: Representation of a neutral current Drell-Yan process in a proton-proton collision.

2.4.1 Drell Yan and W Production

When the final state $A = \gamma^*$, Z_0, we have a *Drell Yan* (DY) process, while when $A = W^\pm$, we refer to the process as *W production*.

These processes are crucial in high energy physics, in fact both the Z and the $W^\pm$ bosons have been discovered in hadronic collisions experiments [11, 12].

We draw the Feynman diagrams of the processes:
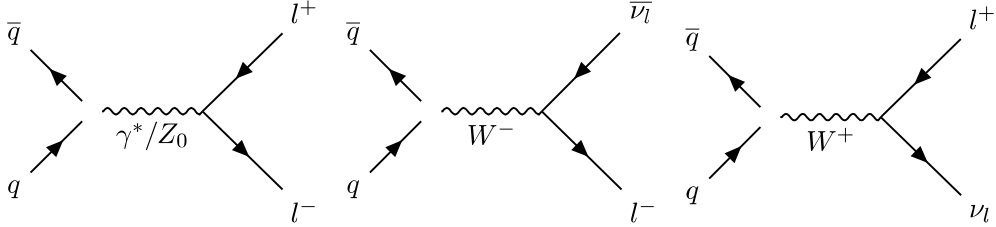
where the first on the left is the Drell Yan process, while the other two involve respectively the W^- and W^+ boson.

As one can expect, the process mediated by the boson Z_0 allows to have more precise measures, since both the particles of the final state are charged and interacting; neutrinos, on the other hand, are not measured in detectors and are only visible in terms of lost momentum. The study of these processes is important for many reasons, such as further investigating the structure of the proton and determining with better precision the quantities associated with the weak interaction bosons. It is of particular interest to determine whether the observed bosons are the one predicted by the Standard Model.

We now want to focus on the process producing W, where the Feynman rule for the interaction vertex is:

$$= \left(\tfrac{ig}{\sqrt{2}}\gamma^{\mu}P_L\right)|V_{ij}|\,\delta_{cc'},$$
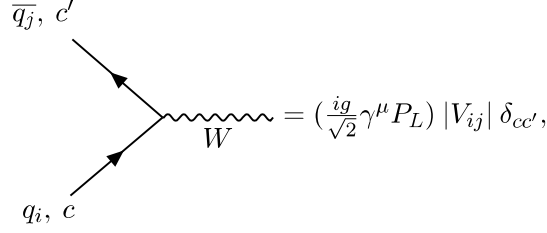
with $P_L = \frac{1-\gamma_5}{2}$, g is the coupling constant of the weak interaction, $|V_{ij}|$ is the element of the Cabibbo-Kobayashi-Mascawa (CKM) matrix, which indicates how likely it is a change of flavour from i to j[1]; lastly, c and c' are the color of the quark and the anti-quark. The presence of P_L stresses how the weak interaction is not chiral. We can also notice that this vertex conserves the color but not the flavour. The same vertex with the boson Z_0 conserves the flavour as well.

The cross section of $q_i\overline{q_j} \to W$ results to be:

$$\hat{\sigma}(q_i\overline{q_j} \to W) = \pi\frac{\sqrt{2}}{3}\,|V_{ij}|^2\,G_F\,M_W^2\,\delta(\hat{s} - M_W^2), \qquad (2.50)$$

where the delta ensures to produce an on-shell W in the final state. The same process involving a virtual boson (W or Z) would be possible but is largely suppressed by the bosons mass, since the propagator expression is $\sim 1/(p^2 - m^2)$.

Using the factorization formula, we can find the cross section of the hadronic process:

$$\sigma(p\overline{p} \to Z) = \sum_{i,j}\int dx_1\int dx_2\,f_i(x_1)\,f_j(x_2)\,\hat{\sigma}(q_i\overline{q_j} \to W) \qquad (2.51)$$

In order to do so, we can introduce the two variables:

$$\tau = \frac{\hat{s}}{s} = x_1 x_2, \qquad (2.52)$$

$$y = \frac{1}{2}\log\left(\frac{p_W^0 + p_W^z}{p_W^0 - p_W^z}\right) = \frac{1}{2}\log\left(\frac{x_1}{x_2}\right) \qquad (2.53)$$

y is the rapidity and it's defined relative to the beam axis z. It is commonly used as a relativistic velocity; this is very convenient since, considering a Lorentz boost along z, the rapidities are additive. Using eq.2.52, eq.2.51 becomes:

$$\sigma(p\overline{p} \to Z) = \sum_{i,j}\pi\frac{\sqrt{2}}{3}\,|V_{ij}|^2\,G_F\,M_W^2\int dx_1\int dx_2\,f_i(x_1)\,f_j(x_2)\,\delta(x_1 x_2 s - M_W^2) =$$

$$\sum_{i,j}\pi\frac{\sqrt{2}}{3}\,|V_{ij}|^2\,G_F\,M_W^2\int dx_1(\frac{\tau}{x_1})\,f_i(x_1)\,f_j(\frac{\tau}{x_1})$$

---

[1] $V_{ud} \sim V_{cs} \sim V_{tb} \sim 1$, while the other elements are ~ 0.

where the last integral is usually called partonic luminosity and it is indicated with $\mathcal{L}_{ij}(\tau)$. If we approximate the quark distribution function with $f_i(x) \sim \frac{1}{x^{1+\delta}}$, where δ is a proper value minor than 1, then:

$$\mathcal{L}(\tau) = \int_\tau^1 \frac{dx}{x} \frac{1}{x^{1+\delta}} \left(\frac{x}{\tau}\right)^{1+\delta} = \frac{1}{\tau^{1+\delta}} \log \frac{1}{\tau}, \tag{2.54}$$

so we obtain that the dependence on s of the cross section is:

$$\sigma_W \sim \tau^{-\delta} \log \frac{1}{\tau} \sim \left(\frac{s}{M_W^2}\right)^\delta \log \left(\frac{s}{M_W^2}\right) \tag{2.55}$$

we have a logarithmic dependence on s, this reflects the fact that, when the energy of the colliding hadrons increases, the quarks can take up a smaller fraction x from the proton momentum and the process still happens. As is shown in fig.2.4 and 2.5, sea quarks have a high probability to take up a *small* fraction of the proton momentum, while this probability vanishes for valence quarks.

2.4.2 Hadronization and Jet Definition

As showed in figure 2.1, the strong interaction coupling constant becomes bigger when the energy lowers, which corresponds to increasing distance between particles. It follows that, for a free parton, is energetically convenient to hadronize, meaning combinig with a particle from the sea. If the initial parton is, for example, a quark, it will combine with a antiquark from a quark-antiquark pair; the remaining quark will have to combine in turn and so on. The result is a 'jet' of hadrons and mesons. It is important to point out that a *jet* is not a elementary object or a clear concept, but actually a tool defined though a algorithm. There also are different used algorithms, and so different *jets* identified in the same final space.

Although there are different ways to define a jet, every algorithm needs to have certain properties:

1. Robustness regarding infrared (IR) and collinear divergences.

2. Facility in both the experimental and theoretical implementation.

3. Non-perturbative corrections have to remain small.

The first attempt at a clustering algorithm for jet definition was formulated by Sterman and Weinberg [30]. In this definition, different particles were part of the same jet if they were soft (their energy had to be $< \epsilon\sqrt{s}$) or collinear (the angle between two particles in the same jet had to be $\theta_{ij} < \delta$). ϵ and δ are the resolution parameters. Nowadays, there also are other possible definition, for example though the k_T algorithm.

Chapter 3

Methodology

3.1   Determination of Parton Distribution Functions

We have seen that PDFs must be extracted from the data. The simplest way of doing this is to choose a model: a fixed functional form that is fitted to the data. This, however, introduces a problem of model dependency, which makes uncertainties particularly difficult to calculate.

In 2002, an alternative was suggested, namely to use *neural networks* for the determination of PDFs [15]. This approach allows for an architecture with a high number of parameters, accommodating virtually any functional form, as well as a method to reliably determine uncertainties.

3.2   NNPDF Method Overview

We now introduce the methodology employed by the NNPDF collaboration to determine PDFs. This methodology has evolved over time; therefore, we will reference only the latest version at the time of writing this thesis, specifically NNPDF4.0.

NNPDF determines PDFs using as an unbiased modeling tool Neural Networks (NN), as we will discuss below. Furthermore, in the NNPDF approach, artificial Monte Carlo pseudodata replicas are generated to ensure reliable error propagation.

Each replica $f^{(r)}$ ($r = 1, 2, ..., N_{rep}$) is equally likely, so the mean value is:

$$< f >= \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} f^{(r)}, \tag{3.1}$$

and the variance is calculated as:

$$Var[f] = \frac{1}{N_{rep}} \sum_{r=1}^{N_{rep}} (f^{(r)} - < f >)^2. \tag{3.2}$$

Figure 3.1 shows this aspect: a total of 200 replicas of the gluon are shown (on the left panel), as well as the correspondent central replica and variance (on the right panel).
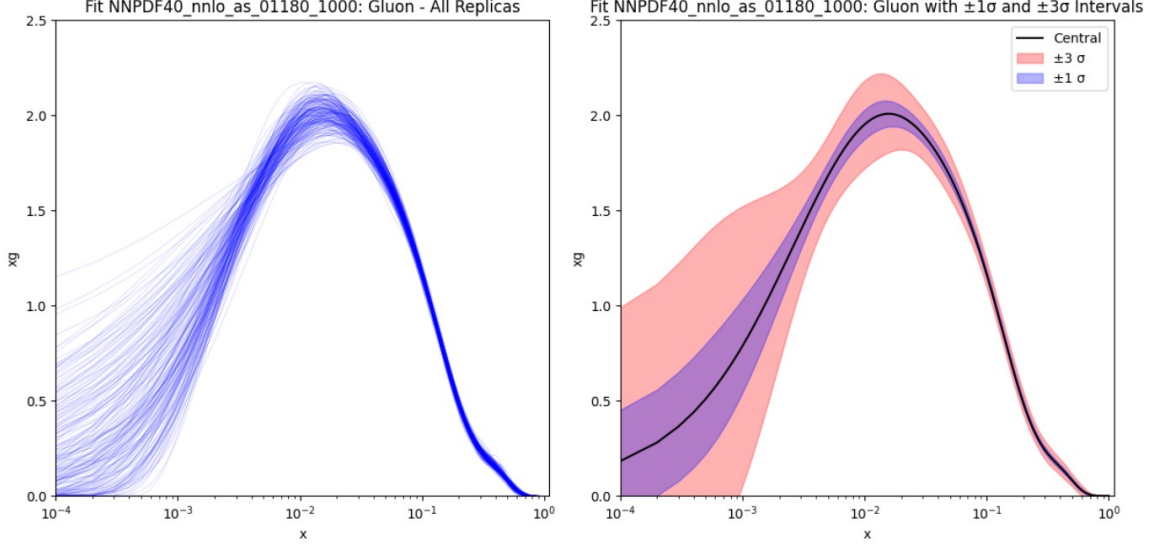

Figure 3.1: Results of the fit NNPDF40_nnlo_as_01180_1000, considering 200 replicas. On the left panel all the replicas are plotted. On the right panel, it is plotted the central replica with intervals corresponding to one standard deviation (1σ, blue) and three standard deviations (3σ, red).

3.2.1 General Structure

We now present the general structure of the NNPDF4.0 fitting framework.
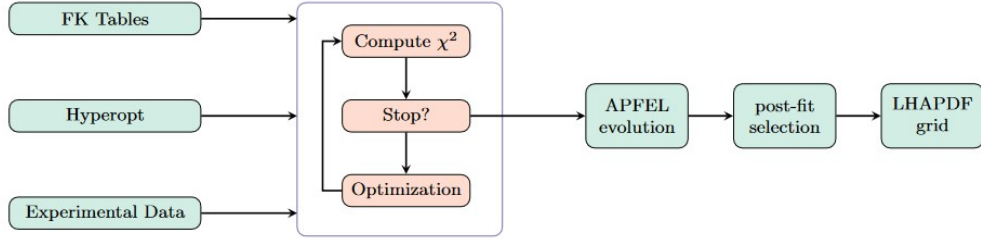
Figure 3.2: Diagrammatic representation of the NNPDF fitting framework. The blue box contains the minimization of the χ^2 figure of merit. On the left there are the three inputs needed to perform the fit.

A good representation of all the main steps needed to obtain a fit are shown in Figure 3.2. The fit requires three main inputs: first, theoretical calculations of physical processes, which are encoded in precomputed tables (FK-tables); second, experimental data, as well as fully correlated uncertainties encoded in a covariance matrix (possibly also including the-

oretical uncertainties); third, hyperparameter settings that determine the particular fitting methodology adopted.

In the blue box of the figure, the core of the computation is represented, which is also showed with more details in Figure 3.3. Here, the neural network optimization algorithm, with settings determined by the hyperoptimization, finds the best fit to data by minimizing a figure of merit (in our case, χ^2). We will explain below what hyperoptimization is, for the moment it is enough to state that it selects the best methodology.

Starting from a matrix of momentum fraction x values, the neural network is trained. The trained NN is then used to construct unnormalized PDFs, which are then normalized according to the constraints we will describe below. After this, we obtain the PDFs at an input scale Q_0.
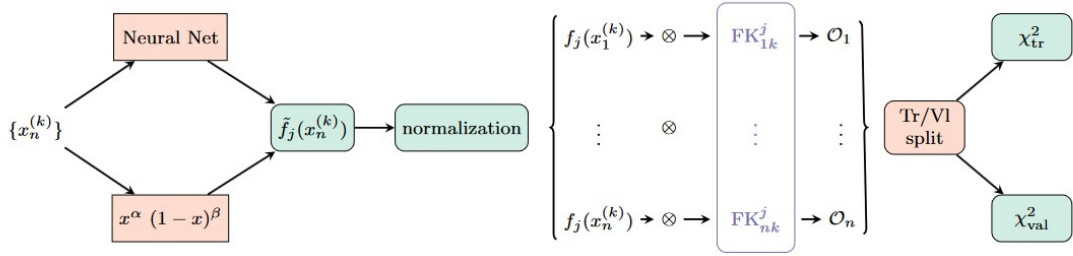

Figure 3.3: Zoom on the calculation of the χ^2 in the NNPDF fitting framework as a function of the values of $\{x_n^{(k)}\}$ for the different datasets. This was previously contained in the blue box of figure 3.2.

3.2.2 Loss Function

We assume that the experimental uncertainties are Gaussian, so a natural choice for the target function is the χ^2, which we define as:

$$\chi^2 = \sum_{i,j=1}^{N_{\text{data}}} (D_i - P_i) \, \text{cov}_{ij}^{-1} \, (D_j - P_j),$$ (3.3)

where D_i are the experimental values of data-point i, P_i the corresponding prediction of the NNPDF model, and cov_{ij} denotes the covariance between the data-points with label i and j. The experimental covariance matrix is obtained as:

$$(\text{cov}_{\text{exp}})_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \left(\sum_{m=1}^{N_{\text{mult}}} \sigma_{i,m}^{(\text{norm})}\sigma_{j,m}^{(\text{norm})} + \sum_{l=1}^{N_{\text{corr}}} \sigma_{i,l}^{(\text{corr})}\sigma_{j,l}^{(\text{corr})}\right) D_i D_j, \quad (3.4)$$

where $\sigma_i^{(\text{uncorr})}$ are the uncorrelated uncertainties, m runs over the N_{norm} multiplicative normalization uncertainties, $\sigma_{i,m}^{(\text{norm})}$, and l runs over the N_{corr} other correlated systematic uncertainties, $\sigma_{i,l}^{(\text{corr})}$.

This value is also referred to as experimental χ^2, since it measures the deviation from the actual experimental data. Although other definition of the χ^2 can be adopted (and are adopted in the NNPDF framework), in this thesis whenever we will enconter a χ^2 value, we will always refer to this definition.

3.3 Neural Network Architecture

In the NNPDF framework, NNs are used to parametrize the PDFs. The use of neural networks allows for a flexible and unbiased parametrization, capable of fitting the data without imposing a predetermined functional form. The methodology is designed to leverage the power of modern deep learning techniques to produce reliable and accurate PDFs.
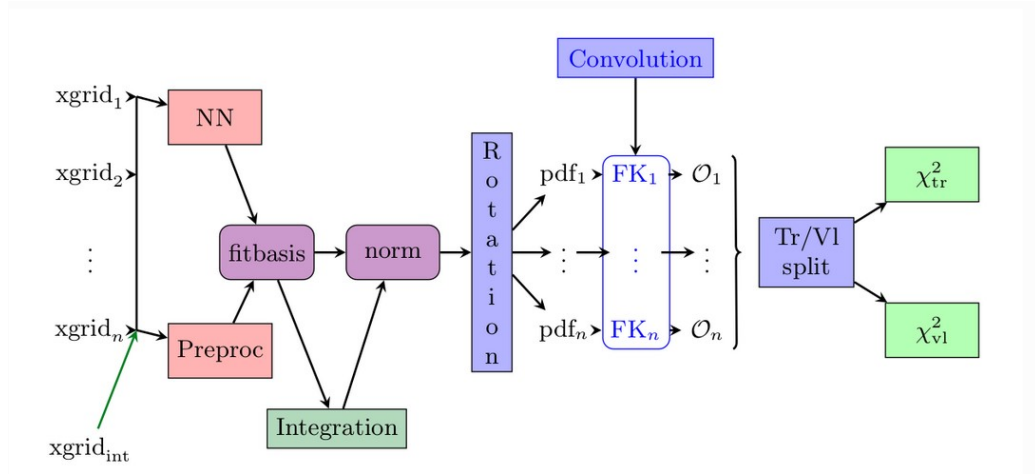
Figure 3.4: Schematic representation of the neural network architecture used in NNPDF4.0.

The architecture of the neural network is crucial for its performance. The NNPDF methodology employs a feed-forward multilayer perceptron neural network. This type of network consists of multiple layers of nodes (neurons), where each node in a layer is connected to every node in the subsequent layer. The network takes the input variable x and transforms it through a series of hidden layers, each applying a non-linear activation function, eventually producing the output PDFs.

A schematic representation is shown in figure 3.5. In the figure, the blue circles correspond to the nodes of the graph and each row is a layer. Here, the input to each activation function corresponds to the set of all outputs of the previous layer as represented by the edges.

At each node of each layer, an *activation function* is computed to obtain its output value, taking as an input all the outputs of the precedent layer with different weights. The output
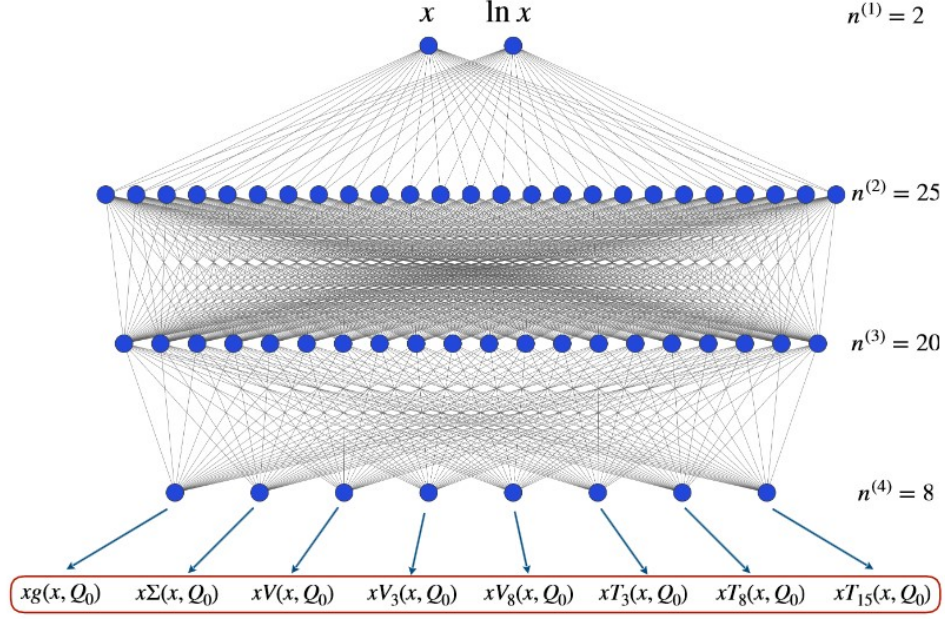
Figure 3.5: Schematic representation of the neural network parametrization used in NNPDF4.0.

of a node i of the layer l is obtained as:

$$\xi_i^{(l)} = g\left(\sum_j w_{ij}^{(l)} \xi_j^{(l-1)} + b_i^l\right), \qquad (3.5)$$

where $g(x)$ is the activation function, $w_{ij}^{(l)}$ are the weights of the different nodes and b_i^l are called biases; these are the free parameters of the neural network. The activation function $g(x)$ can have various forms, provided it is nonlinear and monotonic.

We can notice that the first nodes of the input layer are set as x and $\ln x$: this is because PDFs are believed to scale logarithmically at small x and linearly in the large x region [15].

The architecture can be customized with different numbers of layers, nodes per layer, activation functions, and initialization methods. This flexibility is one of the key strengths of the NNPDF approach.

Training the neural network involves adjusting the parameters (weights and biases) to minimize the loss function, which measures the difference between the prediction and actual data. In the NNPDF methodology, this is done using gradient descent algorithms such as *Adam* or *RMSprop*, which are efficient and well-suited for handling large datasets and complex models.

The parameters defining the model – commonly referred to as the models' hyperparameters – such as the number of layer and the number of nodes per layer, are themselves determined

through a semi-automated procedure, further contributing to a solid and unbiased method.

In Figure 3.4, the architecture of NNPDF4.0 is summarized. The figure displays an integration, which has the scope of fixing the normalization factors, while ensuring the momentum sum rules are respected. These conditions are implemented as a neural network layer which computes the normalization coefficients for each flavour.

To ensure the physical reliability of the PDFs, certain constraints are imposed during the training process. But first, we describe the algorithm used to stop the fitting process.

3.3.1 Stopping Algorithm

Since the NN can fit virtually any functional form, it can easily overfit the data. This means that neural networks will learn noise (such as statistical fluctuations) in the data, rather than the underlying law. In order to avoid overfitting, NNPDF4.0 employs a patience algorithm, depicted in Figure 3.6.

In order to understand how this algorithm is implemented, we consider the outputs of the neural network: PDFs $f_i(x_n^{(k)})$ of flavor i at an input scale Q_0. The outputs of the neural network are convoluted with FK tables encoding the theory calculations and the evolution from the parametrization scale to the scale of the hard process. This convolution provides the corresponding observable $\mathcal{O}_n$.

For hadronic observables, the observables predicted by the NN are obtainable as:

$$\mathcal{O}_n = \text{FK}_{ijkl}^n f_i(x_n^{(k)}, Q_0) f_j(x_n^{(l)}, Q_0), \tag{3.6}$$

while for DIS observables we have:

$$\mathcal{O}_n = \text{FK}_{ik}^n f_i(x_n^{(k)}, Q_0). \tag{3.7}$$

These predicted observables are the ones we compare to the corresponding experimental values to calculate the χ^2 (see Section 3.2.2).

As part of the algorithm we are about to introduce, these observables are randomly separated into a training and a validation set. The first is seen by the NN, while the second is invisible and it is used as a control set. For both the sets, the corresponding training loss χ_{tr}^2 and validation loss χ_{val}^2 are calculated. It is important to notice that, while the experimental χ^2 is computed with the deviation between experimental data-points and NN predictions, the values of χ_{tr}^2, χ_{val}^2 are computed as the deviation between pseudodata $\mathcal{O}_n$ and the NN predictions.

The patience algorithm is based on the look-back cross-validation stopping method [27]. The optimal length of the fit is determined by the absolute minimum of χ_{val}^2 evaluated over a sufficiently large number of iterations of the minimizer. Specifically, the stopping algorithm keeps track of the training step with the lowest χ_{val}^2, and as soon as this value does not
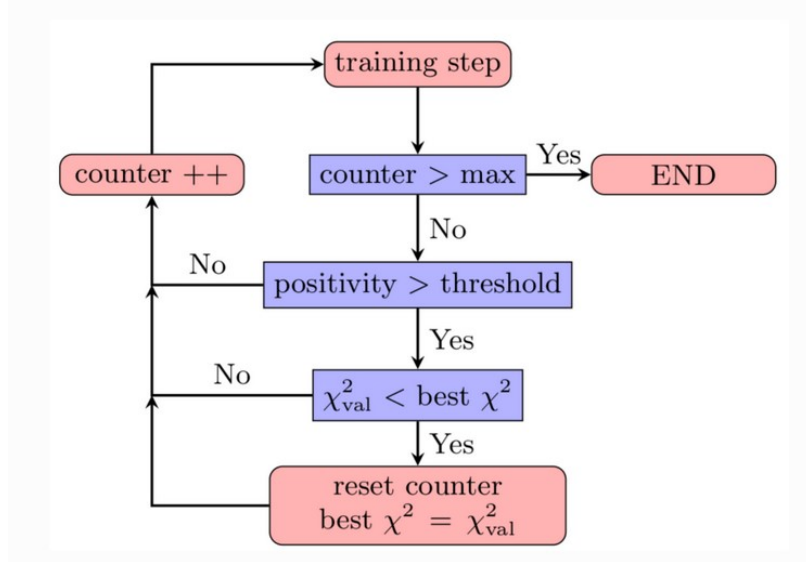
Figure 3.6: Stopping algorithm used in NNPDF. Training stops when validation performance ceases to improve.

improve for a given number of steps (set equal to a percentage of the maximum number of training epochs), the fit is finalized. This is known as early stopping and helps in finding a good balance between fitting the training data and generalizing to unseen data.

3.3.2 Positivity Constraints

The PDFs must be positive for all values of x and Q^2 at Leading Order (LO). This is a necessary condition since PDFs are probability distributions at LO. However, this might not be the case at higher orders, where subtraction schemes are part of the picture and the only strict condition is the positivity of the cross section.

This situation was previously faced allowing negative PDFs to exist, while also adding a penalty term to the loss function for any negative values of the PDFs. However, it has now been shown that that PDFs for individual quark flavors and the gluon in the $\overline{MS}$ factorization scheme are non-negative [25]. For this reason, now the positivity condition on these PDFs is imposed, along with the constraint of positivity of physical cross-sections. The constraint is still obtained by adding a penalty factor to the loss function of the negative PDFs, but the values of these penalty factors are chosen in such a way that the constraint is enforced with sufficient accuracy in all cases.

3.3.3 Integrability Constraints

The PDFs must also satisfy integrability conditions, ensuring that they behave correctly at the boundaries of the x range.

29

In order to fulfil the momentum and valence sum rule, integrability for small values of x is required. We impose respectively:

$$\lim_{x \to 0} x^2 f_k(x, Q) = 0, \tag{3.8}$$

for every value of Q^2 and $f_k = g, \Sigma$, and:

$$\lim_{x \to 0} x f_k(x, Q) = 0, \tag{3.9}$$

for every value of Q^2 and $f_k = V, V_3, V_8$.

A similar condition can also be imposed on $f_k = T_3, T_8$, since both perturbative and unperturbative theory arguments suggest that the first moments of these combinations are finite. So, we extend Eq. 3.9 to $f_k = T_3, T_8$, for every value of Q^2. It has been noticed that the condition in Eq.3.9 is always respected by the fits, while the conditions in Eq.3.8 can be violated, and thus have to be imposed.

The first requirement to meet these condition is to constrain the range of the small-x preprocessing exponents α_i (which we will define in the next section). We supplement the iterative determination of the exponents with the constraints $\alpha_k < 2$ for the singlet and gluon and $\alpha_k < 1$ for the nonsinglet combinations xV, xV_3, xV_8, xT_3 and xT_8. In addition to this, a penalty factor is added to discourage replicas that violate the integrability constraint.

Finally, a post-selection criterion is also adopted to discard all the replicas that don't fit the imposed requirements. In order to achieve this, after the fit is done, the following condition is imposed:

$$\sum_{i=1}^{n_i} \left| x_{int}^{(i)} f_k \left(x_{int}^{(i)} \right) \right| < \frac{1}{2}, \quad f_k = V, V_3, V_8, T_3, T_8, \tag{3.10}$$

where we choose $Q_i^2 = 5 \, \text{GeV}^2$ and in the evolution basis $n_i = 1$ and $x_{\text{int}}^{(1)} = 10^{-9}$, while in the flavor basis $n_i = 3$ and $x_{\text{int}}^{(i)} = 10^{-9}, 10^{-8}, 10^{-7}$.

More information on these constraints can be found in reference [24].

3.4 PDF parametrization

The core problem of PDF determination is the extraction of a continuous function from a discrete set of data. This is, strictly speaking, an ill-defined problem; however, using a proper parametrization, it can be solved. Choosing the parametrization is very important, in fact, a parametrization that is not complex enough introduces a bias in the results. We solve this problem with the use of a NN parametrization, which allows us to work in the limit of an infinite number of parameters, thus making any differentiable function reproducible, as per the universal approximation theorem [17].

In addition to this, it is convenient to introduce preprocessing functions A commonly used form is:

$$xf_i(x, Q_0) = A_i x^{1-\alpha_i}(1-x)^{\beta_i} NN_i(x),$$

(3.11)

where A_i is a normalization constant (determined as described above), α_i and β_i control the small-x and large-x behavior respectively, and $NN_i(x)$ is a single neural network, which gives different outcomes depending on the parton flavor i. The parameters α_i and β_i are randomly sampled from a range that is determined in a self-consistent manner [28, 29]. Specifically, upon making a change to the methodology or dataset, an initial fit is performed for which the effective exponents are calculated for each distribution using:

$$\alpha_{\text{eff},i}(x) = \frac{\log f_i(x)}{\log \frac{1}{x}}, \quad \beta_{\text{eff},i}(x) = \frac{\log f_i(x)}{\log(1-x)}.$$

(3.12)

In later fits, the sampling ranges for the α and β exponents are taken as uniform intervals, twice the size of the 68% confidence interval of the effective exponent. This process is repeated iteratively until the sampling range no longer changes.

The final output of the described methodology is presented in the so-called *evolution basis*, defined as:

$$\Sigma = u + \bar{u} + d + \bar{d} + s + \bar{s} + 2c,$$

(3.13)

$$V = (u - \bar{u}) + (d - \bar{d}) + (s - \bar{s}),$$

(3.14)

$$V_3 = (u - \bar{u}) - (d - \bar{d}),$$

(3.15)

$$V_8 = (u - \bar{u} + d - \bar{d}) - 2(s - \bar{s}),$$

(3.16)

$$T_3 = (u + \bar{u}) - (d + \bar{d}),$$

(3.17)

$$T_8 = (u + \bar{u} + d + \bar{d}) - 2(s + \bar{s}),$$

(3.18)

$$T_{15} = (u + \bar{u} + d + \bar{d} + s + \bar{s}) - 3(c + \bar{c}),$$

(3.19)

$$c^+ = c + \bar{c},$$

(3.20)

$$g = g.$$

(3.21)

This basis is particularly convenient since it decouples many of the evolution equations (see Chapter 2). An alternative to this is the *flavor basis* $f_i = u, \bar{u}, d, \bar{d}, s, \bar{s}, c, g$. However, it has been tested that the resulting PDFs remain largely unchanged upon changes to the choice of parametrization basis [24].

3.5 Hyperoptimization

As anticipated, an important aspect in NNPDF4.0 consists in the determination of the model hyperparameters. Hyperparameters are values that fix the architecture of the neural network,

the training rate, and the specific SGD variant to be used. Basically, hyperoptimization selects the methodology, just like gradient descent selects the values of weights and thresholds of the neural net. More specifically, the automatic hyperoptimization routine uses the improved efficiency provided by Python libraries such as **TensorFlow**. This enables us to test $O(10^3)$ different hyperparameter setups by performing fits with them and ranking the setups through a k-folds cross-validation algorithm.

In order to find the best set of parameters, we look for the combination which best generalize to unseen data, quantifying this with a proper figure of merit L.
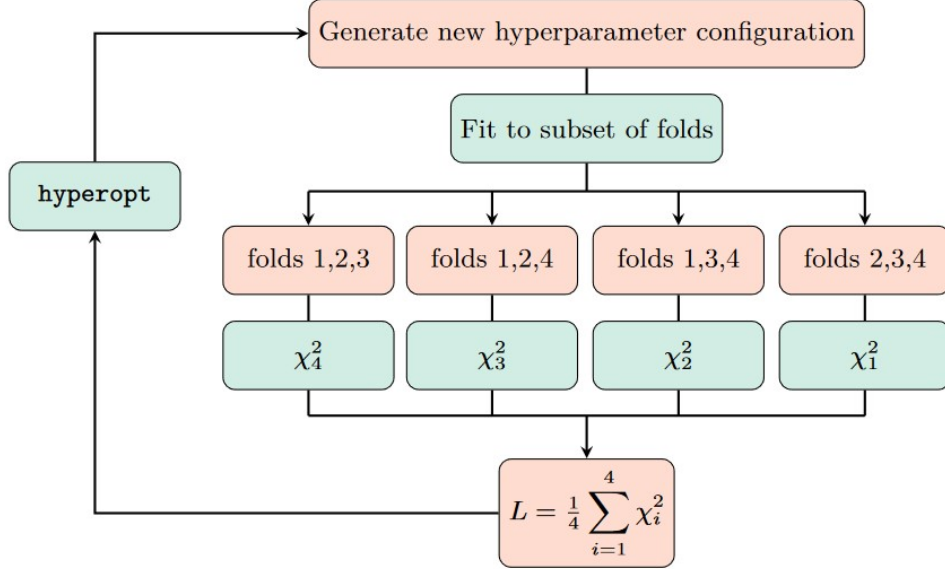

Figure 3.7: Diagrammatic representation of the k-fold algorithm used for the hyperparameter optimization.

The k-folds cross-validation algorithm does not use only a single test set, but instead divides the full dataset into $k = 4$ subsets of data. These subsets are called folds. Each fold is representative of the full dataset both in terms of kinematic range and scattering processes. Then, $k - 1$ of these folds are divided into training and validation datasets that are used to do a fit, while leaving out a k-th fold that will be used as a test set. This is repeated k times, resulting in k fits where for each fit a different fold is used as the test set.

After performing these steps (as in Figure 3.7), we can define the loss function L:

$$L = \frac{1}{k} \sum_{i=1}^{k} \chi_i^2, \tag{3.22}$$

where χ_i^2 is the χ^2 evaluated to the datasets in the i-th fold using the PDF obtained when the i-th fold was left out. The optimal hyperparameter setup is the setup for which L is

minimized.

Further details on the implementation can be found in the NNPDF documentation and relevant literature [4, 13, 14, 15, 24].

Chapter 4

NNPDF4.0 Datasets

4.1   Datasets Introduction

In this chapter, we review the various experiments and processes that contribute to the determination of parton distribution functions (PDFs) in NNPDF4.0.

Figure 4.1 illustrates the kinematic coverage of the approximately 4000 data points used in the fitting procedure, displayed in the $Q^2 - x$ plane, where Q^2 id the trasferred momentum and x is the Bjorken variable (defined in Chapter 2). The distribution of points along the x-axis is crucial for ensuring robust PDFs. As observed, sufficient data coverage is available only for $x > 10^{-4}$, which represents the current small-x boundary. The spread of points along the Q^2-axis is managed using the Altarelli-Parisi splitting functions, which allow the data to be referenced to a common Q_0^2 value.

Different processes constrain PDFs for different parton flavors, emphasizing the importance of diversifying the experimental data sources.

Within the structure of the NNPDF code, these data points are categorized into distinct *datasets*, which group data from the same process and experiment. The datasets vary in size, errors, and their position in the $x - Q^2$ plane, yet each contributes to the overall PDF determination. Some datasets play a more prominent role due to the higher number of data points or smaller errors, while others are strategically important for constraining PDFs at low x or near $x \sim 1$, despite having fewer points or larger uncertainties.
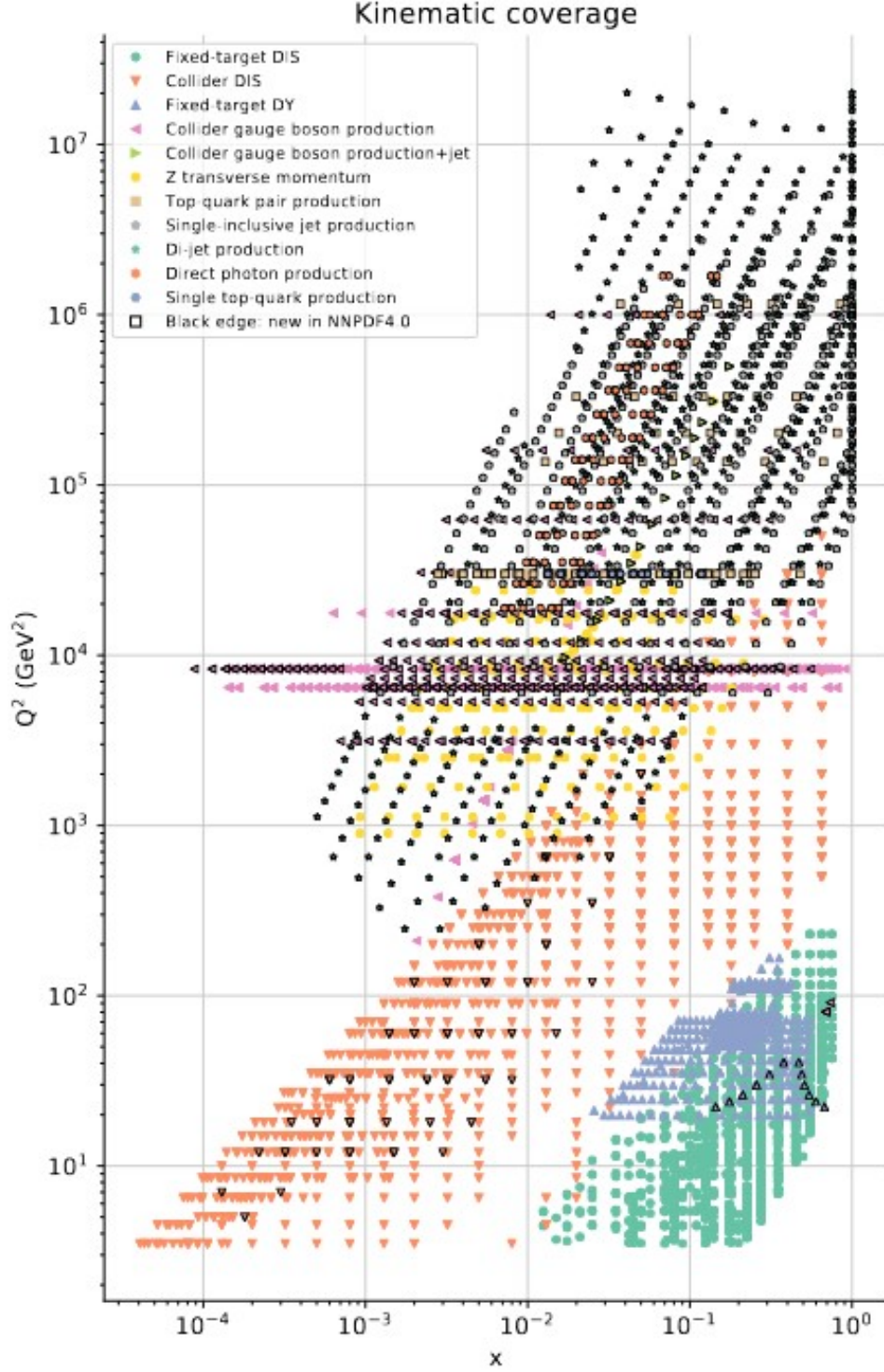
Figure 4.1: Kinematic coverage used for PDF determination in NNPDF4.0, represented in the (x, Q^2) plane. x is the fraction of momentum taken by the parton, and Q^2 is the transferred momentum in the scattering [24].

4.1.1 Deep Inelastic Scattering

Deep Inelastic Scattering (DIS) is a key source of experimental data for PDF determination, providing abundant and precise measurements. As shown in Figure 4.1, DIS data (depicted by light blue points and pink triangles) span a wide range along the x-axis, from $x = 4 \times 10^{-5}$ to $x = 0.750$, and contribute approximately 3000 data points. Along the Q^2 axis, DIS data are confined to the range $1.0 < Q^2 < 5 \times 10^4$ GeV^2.

DIS processes are divided into Neutral Current (NC) and Charged Current (CC) interactions, mediated by photons (γ), Z_0 bosons, and $W^\pm$ bosons, respectively. Together, NC and CC processes help determine the light quark flavors. Some jet data also originate from DIS.

The observables from DIS, such as ratios, cross-sections, and form factors, are measured in several experiments, including NMC, SLAC, BCDMS, CHORUS, NuTeV, HERA, NOMAD, and EMC.

4.1.2 Drell-Yan and $Z/W^\pm$ Production

The Drell-Yan (DY) process is another cornerstone for PDF determination, as discussed in Chapter 2. The dataset includes approximately 300 fixed-target DY data points, represented by blue triangles in Figure 4.1. DY data, along with DIS data, are crucial for determining the gluon PDF from scale dependence and higher-order corrections. The transverse momentum (p_T) distribution in electroweak (EW) processes also provides insights into the small-x behavior of gluons at leading order (LO).

When the parton momentum fractions, x_1 and x_2, are such that the energy of the partonic subprocess $\hat{s}$ is equal to the mass of a Z_0 or $W^\pm$ boson:

$$\hat{s} = x_1 x_2 \, s = m^2_{W/Z}, \qquad (4.1)$$

a resonance occurs, resulting in $Z/W^\pm$ production. This configuration provides an additional 500 data points from CMS and ATLAS experiments at the LHC.

The fixed-target DY data cover the region $5 \times 10^{-2} < x < 0.5$, $20 < Q^2 < 200$ GeV^2, while collider boson production data cover $10^{-4} < x < 1$ and $Q^2 \sim 10^4$ GeV^2.

4.1.3 Jet Production

Jet production occurs whenever the final state is purely strong-interacting. The process can involve either single-inclusive jet production (one jet in the final state) or di-jet production (two jets in the final state).

Jet data are also essential for PDF determination, particularly the charm jet data from CMS, which are used in conjunction with W production measurements to achieve next-to-leading order (NLO) accuracy [24].
Single-inclusive and di-jet measurements are made at $\sqrt{s} = 8\,GeV$ and $\sqrt{s} = 7,\,8\,GeV$,

respectively. In the case of the single-inclusive jet, the data cover the range $70 \leq p_T^{jet} \leq 2.5 TeV$ and $\mid y^{jet} \mid \leq 3.0$, where p_T^{jet} is the transfered momentum and y^{jet} is the rapidity. For di-jet, on the other hand, the corresponding ranges are $260 \leq m_{jj} \leq 4.27\,TeV$ and $0.0 \leq y^* \leq 3.0$, where m_{jj} is the invariant mass and y^* is he absolute difference of the rapidities of the two jets.

4.1.4   Top Pair and Single Top Production

Top pair production occurs through quark-antiquark annihilation or gluon-gluon fusion, measurable at the LHC. Various processes can also lead to single top quark production. Differential and total cross-sections for top pair production are measured at $\sqrt{s} = 8$ TeV, while single top production is measured at $\sqrt{s} = 7$, 8, and 13 GeV. For single top production, the rapidity ranges are $|y_{t,\bar{t}}| < 3.0$ and $|y_{t,\bar{t}}| < 2.2$ at $\sqrt{s} = 7$ and 8 TeV, respectively. Data are obtained from both CMS and ATLAS.

4.1.5   Direct Photon Production

Isolated photon production is measured at ATLAS at $\sqrt{s} = 8$ and 13 TeV. The measurements provide cross-sections differential in the photon transverse energy E_T^γ across different bins of photon pseudorapidity η_γ. These data, recently added to the dataset, appear to have a mild impact on the gluon PDF at intermediate x values.

4.1.6   Nuclear Datasets

A significant portion of the dataset (approximately 1400 points) originates from deep inelastic or hadronic scattering on deuterium or heavy nuclear targets, such as iron, lead, or copper nuclei ($^{56}_{26}Fe$, $^{208}_{82}Pb$, $^{64}_{32}Cu$). The data are collected from experiments such as NMC, SLAC, BCDMS, E866, E906, E605, CHORUS, and NuTeV.

Nuclear datasets introduce additional complexities due to the presence of the nuclear medium, which modifies the PDFs compared to those of free protons and neutrons. These modifications arise from several factors, including nuclear binding, Fermi motion, and shadowing effects at small x, as well as anti-shadowing and EMC effects at intermediate and large x [35, 36].

In the NNPDF4.0 analysis, nuclear effects are carefully accounted for by incorporating nuclear corrections into the theoretical predictions for processes involving nuclear targets.

Moreover, nuclear datasets provide valuable information on the flavor separation of PDFs, particularly in the case of deuterium targets, which are sensitive to the difference between up and down quark distributions [37]. Data from heavy nuclei, on the other hand, are crucial for understanding the gluon distribution at small x and its evolution with Q^2, as nuclear shadowing becomes significant in this region [38].

4.2 Complete List of the Datasets

In the next table, we list all the dataset used in NNPDF4.0 [24]; we will recall these datasets throughout the thesis using this naming convention.

In the 'Description' column, we report labels to identify the type of process. Some labels are self-explanatory (DIS, DY, JET, DIJET), others require a legend which is provided at the end of the chapter.

Dataset	Description	N data
ATLAS_1JET_8TEV_R06_DEC	JET	171
ATLAS_2JET_7TEV_R06	DIJET	90
ATLAS_DY_2D_8TEV_LOWMASS	EWK_RAP	84
ATLAS_SINGLETOP_TCH_DIFF_7TEV_T_RAP_NORM	HQP_YQ	3
ATLAS_SINGLETOP_TCH_DIFF_7TEV_TBAR_RAP_NORM	HQP_YQ	3
ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM	HQP_YQ	3
ATLAS_SINGLETOP_TCH_DIFF_8TEV_TBAR_RAP_NORM	HQP_YQ	3
ATLAS_SINGLETOP_TCH_R_13TEV	INC	1
ATLAS_SINGLETOP_TCH_R_7TEV	INC	1
ATLAS_TOPDIFF_DILEPT_8TEV_TTRAPNORM	HQP_YQQ	5
ATLAS_TTBARTOT_13TEV_FULLLUMI	INC	1
ATLAS_TTB_DIFF_8TEV_LJ_TRAPNORM	HQP_YQ	5
ATLAS_TTB_DIFF_8TEV_LJ_TTRAPNORM	HQP_YQQ	5
ATLAS_WM_JET_8TEV_PT	EWJ_PT	16
ATLAS_WP_JET_8TEV_PT	EWJ_PT	16
ATLAS_WZ_TOT_13TEV	INC	3
ATLAS_ZHIGHMASS49FB	EWK_MLL	13
ATLASDY2D8TEV	EWK_RAP	48
ATLASLOMASSDY11EXT	EWK_MLL	6
ATLASPHT15_SF	PHT	53
ATLASWZRAP11CC	EWK_RAP	46
ATLASWZRAP11CF	EWK_RAP	15
ATLASWZRAP36PB	EWK_RAP	30
ATLASZPT8TEVMDIST	EWK_PT	64
ATLASZPT8TEVYDIST	EWK_PTRAP	120
ATLASTTBARTOT7TEV	INC	1
ATLASTTBARTOT8TEV	INC	1
BCDMSD_dw_ite	DIS_F2D	254
BCDMSP_dwsh	DIS_F2P	351

Dataset	Description	N data
CDFZRAP_NEW	EWK_RAP	28
CHORUSNBPb_dw_ite	DIS_SNB_Pb	607
CHORUSNUPb_dw_ite	DIS_SNU_Pb	607
CMS_1JET_8TEV	JET	239
CMS_2JET_7TEV	DIJET	54
CMS_SINGLETOP_TCH_R_13TEV	INC	1
CMS_SINGLETOP_TCH_R_8TEV	INC	1
CMS_SINGLETOP_TCH_TOT_7TEV	INC	1
CMS_TTB_DIFF_13TEV_2016_2L_TRAP	HQP_YQ	10
CMS_TTB_DIFF_13TEV_2016_LJ_TRAP	HQP_YQ	11
CMSDY2D11	EWK_RAP	132
CMSWEASY840PB	EWK_RAP_ASY	11
CMSWMASY47FB	EWK_RAP_ASY	11
CMSWMU8TEV	EWK_RAP	22
CMSZDIFF12	EWK_PTRAP	50
CMSTOPDIFF8TEVTTRAPNORM	HQP_YQQ	10
CMSTTBARTOT13TEV	INC	1
CMSTTBARTOT5TEV	INC	1
CMSTTBARTOT7TEV	INC	1
CMSTTBARTOT8TEV	INC	1
D0WMASY	EWK_RAP_ASY	10
D0ZRAP_40	EWK_RAP	28
DYE605_dw_ite	DYP_E605	119
DYE886P	DYP_E886P	184
DYE886R_dw_ite	DYP_E886R	15
DYE906R_dw_ite	DYP_E906R	6
HERACOMB_CCEM	DIS_CCE	42
HERACOMB_CCEP	DIS_CCP	39
HERACOMB_NCEM	DIS_NCE	159
HERACOMB_NCEP460	DIS_NCP	209
HERACOMB_NCEP575	DIS_NCP	260
HERACOMB_NCEP820	DIS_NCP	112
HERACOMB_NCEP920	DIS_NCP	485
HERACOMB_SIGMARED_B	DIS_NCE_BT	27
HERACOMB_SIGMARED_C	DIS_NCP_CH	52
LHCB_WZMU7TEV	EWK_RAP	33

Dataset	Description	N data
LHCB_WZMU8TEV	EWK_RAP	34
LHCB_Z_13TEV_DIELECTRON	EWK_RAP	17
LHCB_Z_13TEV_DIMUON	EWK_RAP	18
LHCBZ940PB	EWK_RAP	9
LHCBZEE2FB_40	EWK_RAP	17
NMC	DIS_NCE	292
NMCPD_dw_ite	DIS_F2R	260
NTVNBDMNFe_dw_ite	DIS_DM_NB	45
NTVNUDMNFe_dw_ite	DIS_DM_NU	45
SLACD_dw_ite	DIS_F2D	211
SLACP_dwsh	DIS_F2P	211

Table 4.1: Table containing all the datasets used in NNPDF4.0, as well as a description of the data contained in each dataset (the legend is in the text), as well as the number of points of each dataset.

The labels in the column 'Description' of Table 4.1 need to be explained. They are formed by two abbreviations divided by an underscore.

The first part of the labels refer to the type of process: DY for Drell Yan, DIS for Deep Inelastic Scattering, EWK for weak boson production, EWJ for W production with one jet, PHT for photon production, INC for inclusive cross section, HQP for single top and top pair production, JET and DIJET as described previously.

The second part of the label, when present, says which observable is measured in that dataset. For nuclear correction datasets, the label is composed by either 'DIS' or 'DY', followed by details on the target or experiment.

Chapter 5

Results

5.1 Defining the problem

The starting point of this thesis can be effectively explained with the plot of Figure 5.1. In this plot, we show a selection of replicas; they were obtained with NNPDF4.0 in a fit with a total of 1000 replicas. In this plot, as in the following ones in this thesis, we fix $Q^2 = 1.65 GeV$.

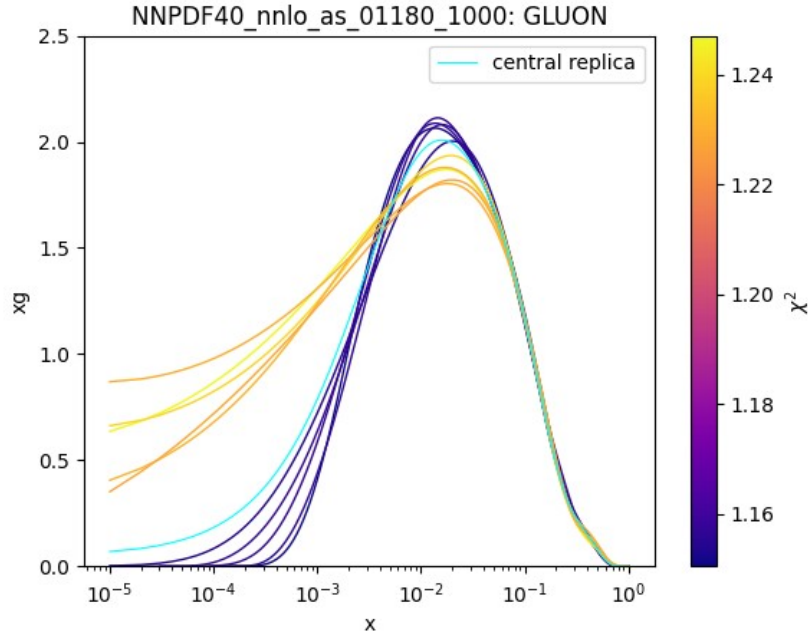Figure 5.1: Guon PDF replicas from a fit named 'NNPDF40_nnlo_as_01180_1000' of 1000 replicas. We display the five replicas with higher and lower χ^2, where the χ^2 value is represented by the color of the replica. The central replica is also represented.

The figure displays the five replicas with the highest experimental χ^2 (defined in Section

3.2.2) values and the five with the lowest, alongside the central replica. Several important observations can be made from this plot.

First, there is a noticeable pattern—replicas with higher χ^2 tend to exhibit a lower peak and a higher tail, whereas those with lower χ^2 show a higher peak and flatter tail.

Second, the central replica (in cyan) does not align with the replicas that have the lowest χ^2, but instead has a slightly lower peak and a more pronounced tail. As discussed in Chapter 3, each replica is independent and equally likely, meaning the central replica represents the mean of all replicas. Therefore, the central replica is not necessarily expected to have the best agreement with the data.

This issue was first highlighted in [32], where the authors exhibited PDFs with χ^2 values lower than that of the central replica and argued that this suggests that the central replica might be subject to sampling biases. However, the fact that the central replica does not correspond to the best χ^2 is actually a positive indication—it suggests that the central replica interpolates the data without overfitting. The distribution of χ^2 values for the replicas forms an approximately Gaussian shape, as shown in Figure 5.2, with the central replica located in the tail of this distribution.
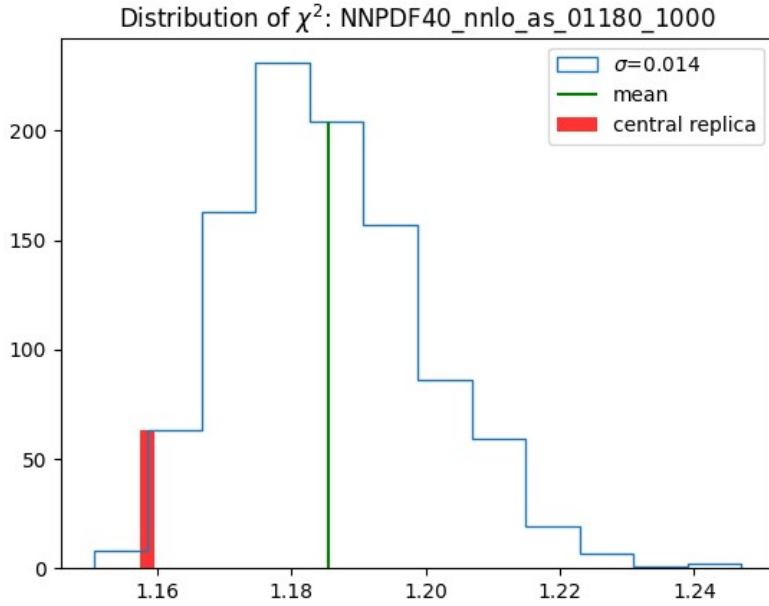

Figure 5.2: Distribution of χ^2 values for the 1000 replicas of the fit. The mean, standard deviation, and χ^2 of the central replica are indicated.

It is important to note that outliers in the χ^2 distribution, including those with the lowest χ^2, are typically good fits to unlikely fluctuations in the data [33]. The absolute minimum χ^2 does not necessarily correspond to the best solution but may instead reflect overfitting.

However, established that these results may be valid and free from sampling biases (as proved by closure tests and future tests [33]), the underlying question remains: why do the replicas with the lowest χ^2 exhibit these distinct features, such as a higher peak and lower tail? Moreover, if these replicas are fitting outliers in the data, can we identify the source of this behavior within the datasets?

In essence, this thesis aims to explore the possibility of applying explainable AI to the NNPDF framework, which could help in understanding these peculiarities and in identifying the different properties of the datasets.

5.1.1 Focusing on the Gluon PDF

This thesis primarily focuses on the gluon PDF. This is because the gluon PDF displays the most strong correlation between specific features (height of the peak) and agreement to the data. In fact, as seen in Figures 5.4 and 5.3, repeating the analysis for sea and valence PDFs reveals much less separation between the best and worst χ^2 replicas.
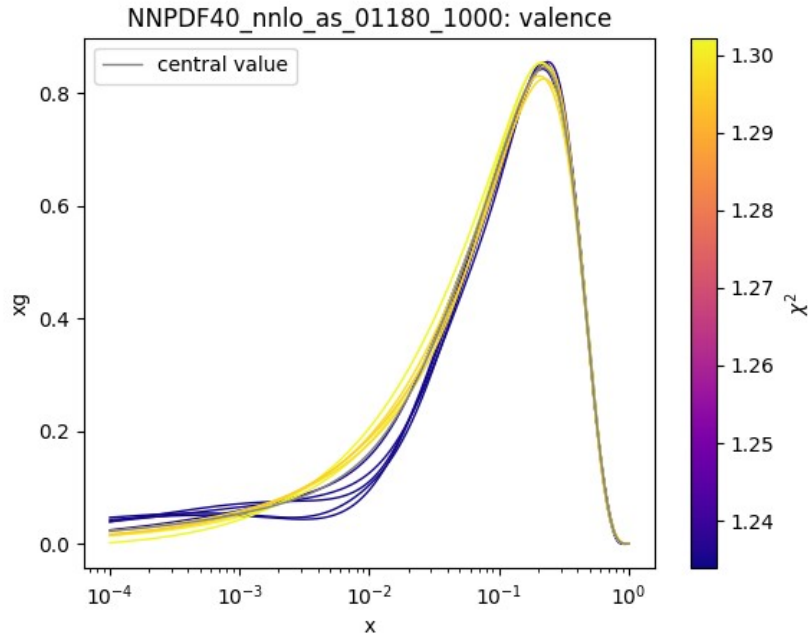

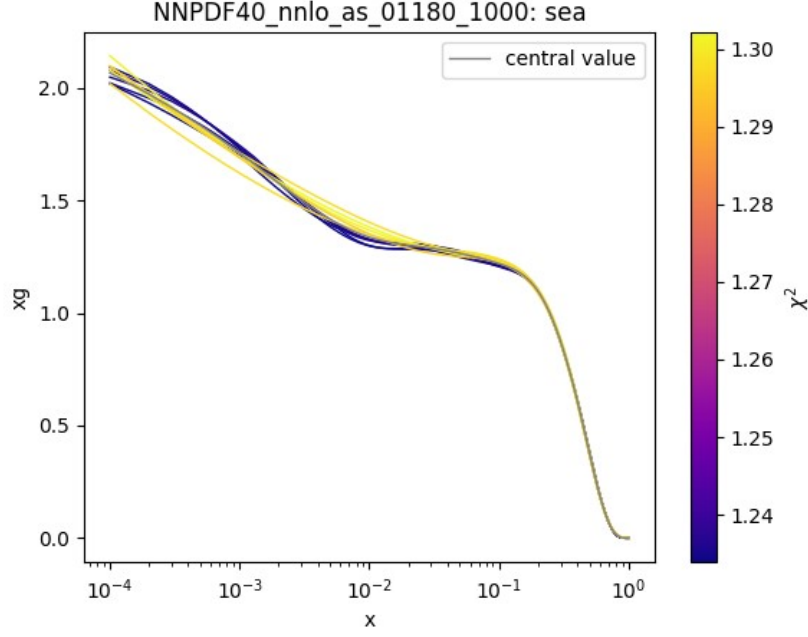Figure 5.3: Same as in Figure 5.1, but for the valence PDF.

45

Figure 5.4: Same as in Figure 5.1, but for the sea PDF.

5.2 K-Folding

A reasonable conjecture, is that the outlier replicas do not generalize. Generalization is optimized using k-folding, which removes data so that the fit quality to the not fitted data can be checked. As described in Chapter 3, this has been used for hyperparameters determination. In order to check wheather these effects follow from the k-folding used for hyperoptimization, we use the same folds.

So, we use four folds which have a balanced dataset division both in terms of the number of data points and their distribution across the x-axis.

As the first step, we replicate the plot shown in Figure 5.1, excluding one fold at a time, resulting in four new fits. A few preliminary remarks are necessary before we delve into the results: first, these fits contain approximately 200 replicas—enough to extract qualitative information but significantly fewer than the 1000 replicas in Figure 5.1. Second, the χ^2 values in these plots refer to the *in-sample* datasets, meaning the χ^2 is not calculated for datasets that are part of the excluded fold.

Throughout this thesis, we will denote a fit with the name of the *excluded* fold or datasets. For example, a fit labeled 'fold 1' indicates that the fit was performed excluding all datasets belonging to fold 1.
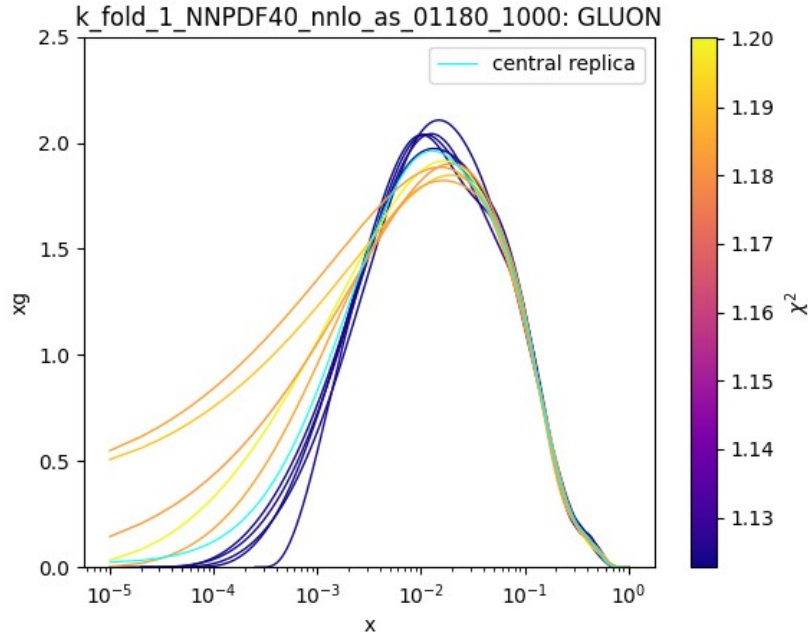
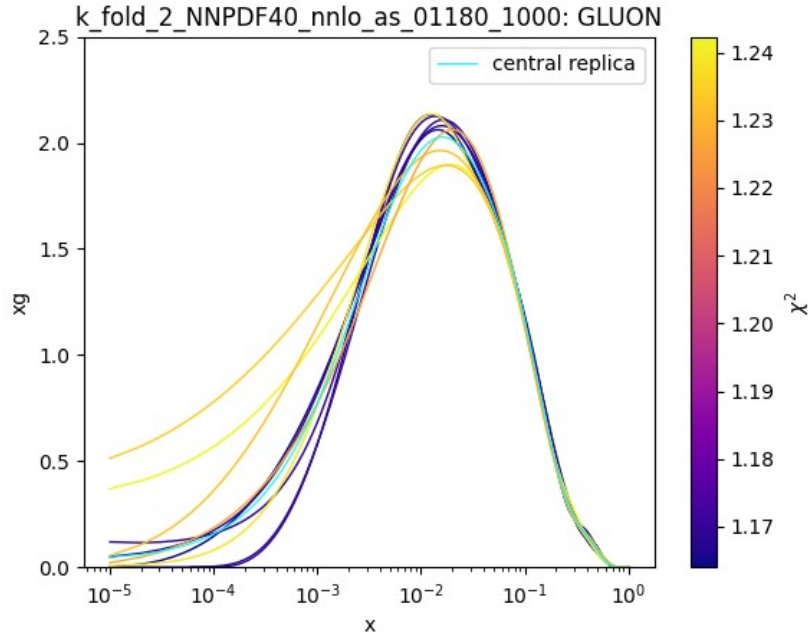Figure 5.5: Same as for Figure 5.1 but excluding the fold in the fit name. We recall that $Q^2 = 1.65\,GeV$.

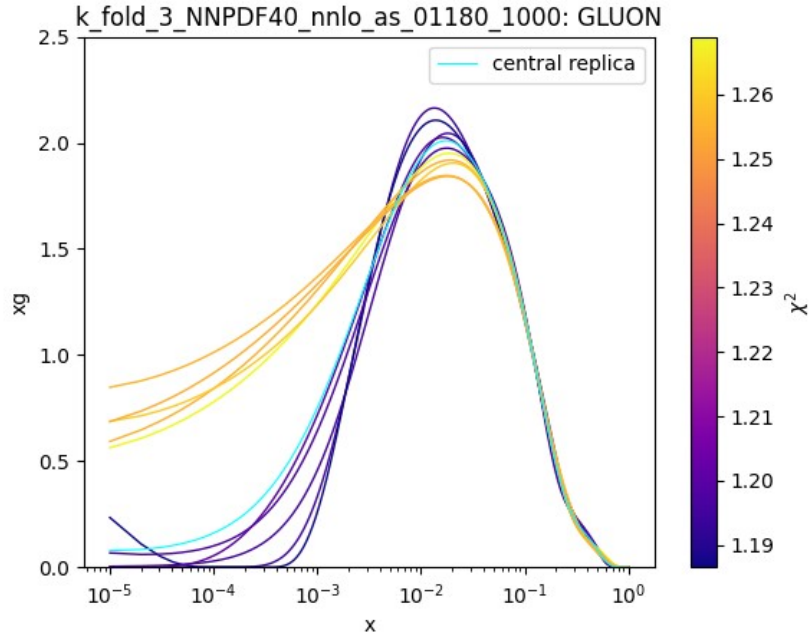Figure 5.6: Same as for Figure 5.1 but excluding the fold in the fit name. We recall that $Q^2 = 1.65\,GeV$.

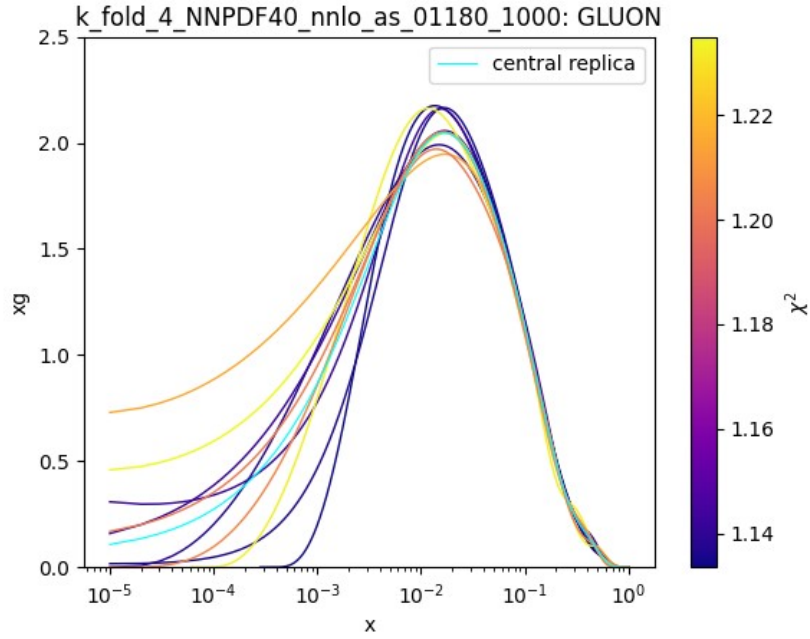Figure 5.7: Same as for Figure 5.1 but excluding the fold in the fit name. We recall that $Q^2 = 1.65 \, GeV$.

Figure 5.8: Same as for Figure 5.1 but excluding the fold in the fit name. We recall that $Q^2 = 1.65 \, GeV$.

The results are shown in Figure 5.5, 5.6, 5.7 and 5.8. We can clearly notice differences between these fits.

When fold 1 is excluded, there is a noticeable separation between the best and worst χ^2 replicas. However, this separation is less pronounced compared to the reference fit. Additionally, the worst χ^2 replicas show considerable dispersion. Among the best χ^2 replicas, one exhibits a steep and rapid change in direction; this kind of rapid change usually suggests overfitting.

When excluding fold 2, results show a weaker separation between the best and worst χ^2 replicas, accompanied by increased dispersion among the higher χ^2 replicas.

When fold 3 is the one out of sample, we observe a clear division similar to that in Figure 5.1, though with slightly higher χ^2 values.

When fold 4 is left out, the division between best and worst χ^2 replicas becomes less distinct, and both sets of replicas are highly dispersed.

We also obtained similar plots by selecting the five replicas with the lowest χ^2 and the five with the highest χ^2 from the excluded datasets. It was hypothesized that the best replicas might exhibit increased χ^2 values on the excluded datasets due to overfitting characteristics. However, as shown in Figures 5.9-5.12, this does not seem to occur, except marginally when excluding fold 1.
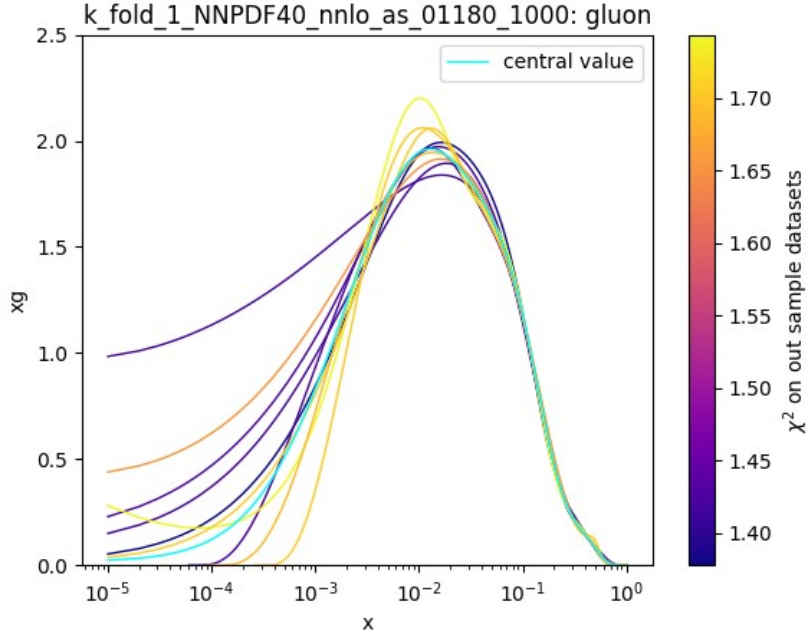

Figure 5.9: Same as in Figure 5.1, but computing the χ^2 on the excluded datasets.
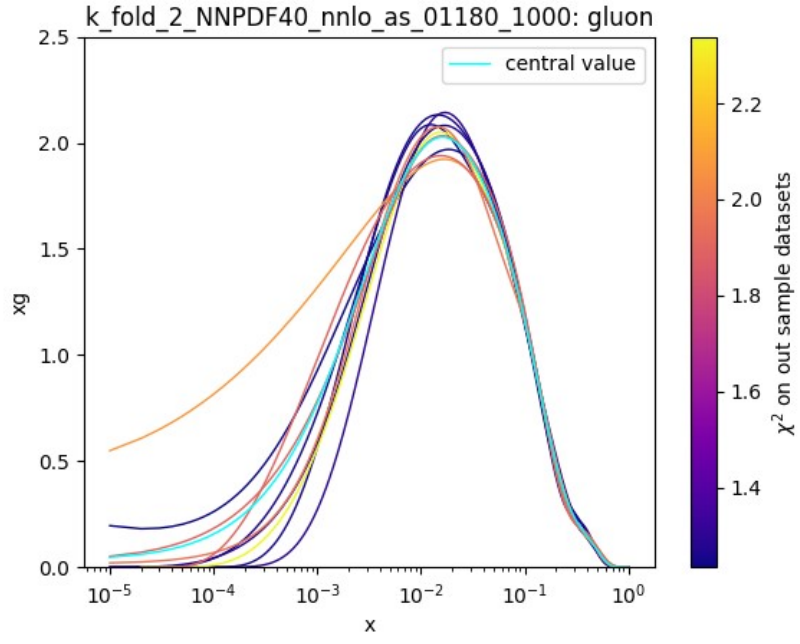
Figure 5.10: Same as in Figure 5.1, but computing the χ^2 on the excluded datasets.
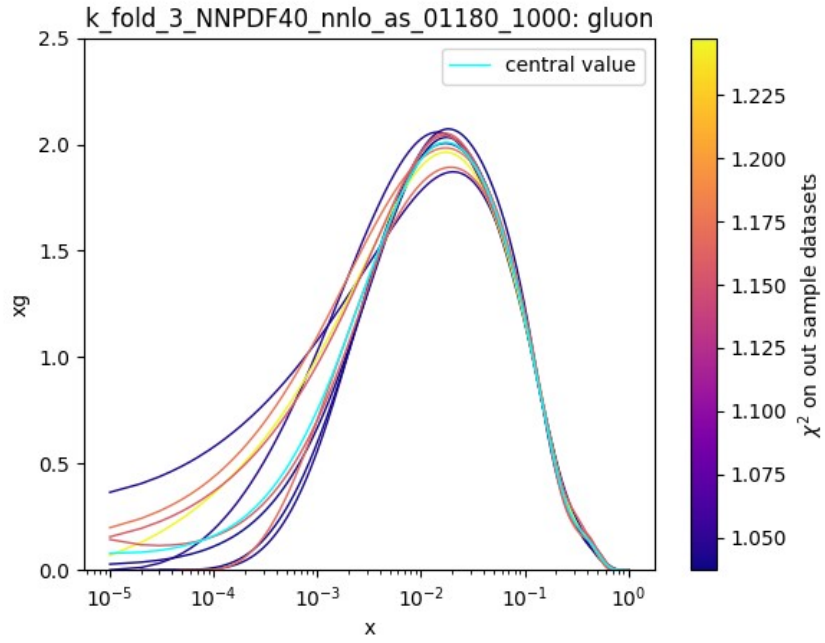

Figure 5.11: Same as in Figure 5.1, but computing the χ^2 on the excluded datasets.
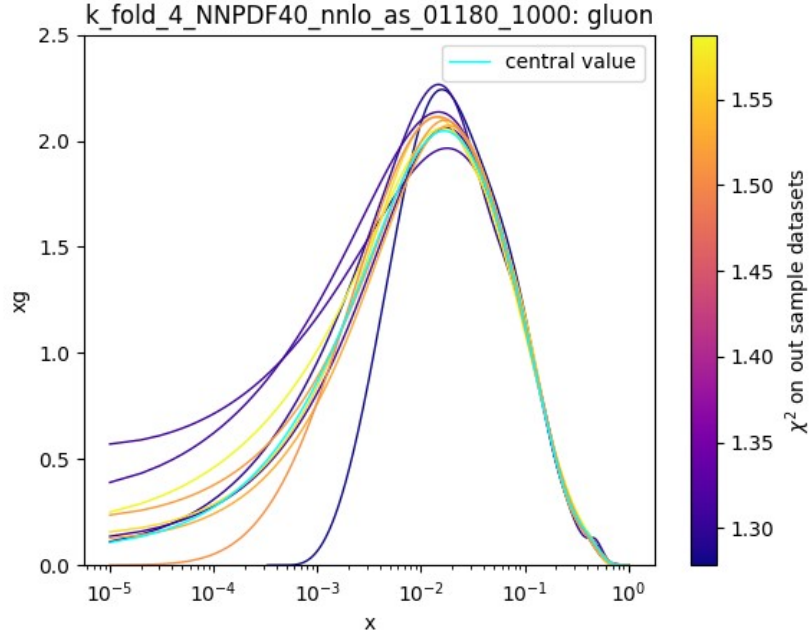
Figure 5.12: Same as in Figure 5.1, but computing the χ^2 on the excluded datasets.

Since the plots reveal that replicas with higher χ^2 on the excluded folds are not necessarily those with lower χ^2 on the included datasets (and vice versa), we aim to compare the χ^2 values on the included datasets with those on the excluded datasets for each replica. To facilitate this comparison, Figure 5.13 presents a scatter plot with four panels. Each panel displays a scatter plot where the χ^2 values from the included datasets are plotted on the x-axis and the χ^2 values from the excluded datasets on the y-axis, with each point representing a replica.

The scatter plots reveal a low correlation between the χ^2 values on the included and excluded datasets across all four panels. This lack of correlation indicates that replicas which perform better on the included datasets do not necessarily perform worse on the excluded datasets.

This further suggests that the method is robust and does not exhibit evidence of overfitting. However, it also shows that the conjecture was not correct: the reason why we have outliers replicas is not that they overfit the data and/or that they do not generalize.

To sum up, excluding fold 1 or fold 3 results in a clearer separation between the best and worst χ^2 replicas, whereas the distinction between these groups is less pronounced when excluding fold 2 or fold 4. However, the analysis of χ^2 values on the excluded datasets shows no significant separation, except for a slight indication when excluding fold 1. This additional analysis confirms the absence of correlation between included and excluded χ^2 values, supporting the robustness of the method.
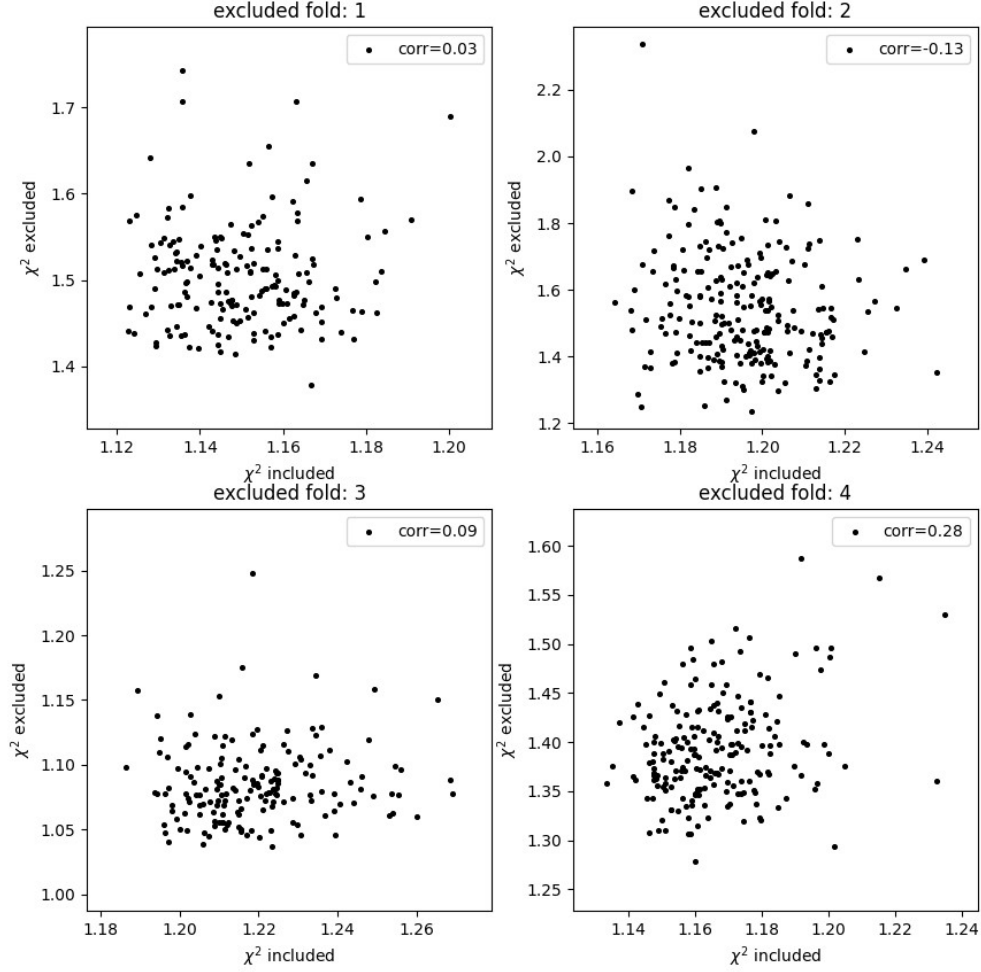
51

Figure 5.13: This figure presents four panels: in each, results from a fit obtained excluding one of the fold are shown. Each point represents a replica, the x-axis value is the χ^2 of the replica on the included datasets, while the y-axis value is the χ^2 of the replica on the excluded fold. In each panel, the correlation between the two axis is also reported.

5.2.1 Fold Composition

As said, the results suggest the absence of overfitting and thus disprove our initial conjecture. Our next conjecture is that the replicas may be outliers *within* folds, thus fitting only some data better. To investigate this, we need to know how the composition of the folds.

In this section, we examine the composition of the folds, which were designed to be balanced along the x-axis and to include various types of processes. The datasets included in each fold are detailed in Tables 5.1 and 5.2. For definitions of the datasets, refer to Table

4.1).

Fold 1	Fold 2
ATLAS_2JET_7TEV_R06	ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM
ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM	ATLAS_WP_JET_8TEV_PT
ATLASPHT15_SF	ATLASTTBARTOT7TEV
ATLASZPT8TEVMDIST	ATLASWZRAP11CC
CMS_SINGLETOP_TCH_TOT_7TEV	ATLASZHIGHMASS49FB
CMSTTBARTOT7TEV	CDFZRAP_NEW
CMSWEASY840PB	CMS_TTBAR_2D_DIFF_MTT_TRAP_NORM
DYE886P	CMSTTBARTOT7TEV
HERACOMB_SIGMARED_B	CMSWMASY47FB
HERACOMBCCEM	D0WMASY
HERACOMBNCEP460	DYE886R_dw_ite
LHCBZEE2FB_40	HERACOMBCCEP
NMC	HERACOMBNCEP575
NTVNBDMNFe_dw_ite	LHCB_Z_13TEV_DIELECTRON
	LHCBWZMU7TEV
	NMCPD_dw_ite
	NTVNUDMNFe_dw_ite
	HERACOMB_SIGMARED_B

Table 5.1: Datasets in Fold 1 and Fold 2

Fold 3	Fold 4
ATLAS_SINGLETOP_TCH_DIFF_7TEV_T_RAP_-NORM	ATLAS_SINGLETOP_TCH_DIFF_7TEV_TBAR_RAP_-NORM
ATLAS_SINGLETOP_TCH_R_13TEV	ATLAS_SINGLETOP_TCH_R_7TEV
ATLAS_WM_JET_8TEV_PT	ATLASPHT15_SF
ATLASLOMASSDY11EXT	ATLASWZRAP36PB
ATLASWZRAP11CF	BCDMSP_dwsh
ATLASZPT8TEVYDIST	CHORUSNUPb_dw_ite
BCDMSD_dw_ite	CMS_2JET_7TEV
CHORUSNBPb_dw_ite	CMS_SINGLETOP_TCH_R_8TEV
CMS_2JET_7TEV	CMSDY2D11
CMS_SINGLETOP_TCH_R_13TEV	CMSTOPDIFF8TEVTTRAPNORM
CMSWMU8TEV	CMSZDIFF12
D0ZRAP_40	DYE605_dw_ite
HERACOMBNCEP820	HERACOMBNCEP920
LHCB_Z_13TEV_DIMUON	LHCBZ940PB
LHCBWZMU8TEV	

Table 5.2: Datasets in Fold 3 and Fold 4

In Figure 5.14, we present visually the composition of the folds: each dataset is color-coded according to the type of process it contains, with the labels described at the end of Chapter 4). The datasets are displayed on the $x-$axis, where the separation between different folds is marked by a dotted line; datasets in the fifth group are not part of any fold and they are always included in our fits. This is because these datasets were not used for the hyperoptimization, so they are not part of the folds, but since they are now available we choose to include them.

On the y-axis, the χ^2 value for each dataset is shown. The χ^2 computation has been made for each replica of the reference fit, and the χ^2 value of each replica is represented by a point.

53

We now want to investigate how the plot in Figure 5.14 changes if we exclude one of the folds.
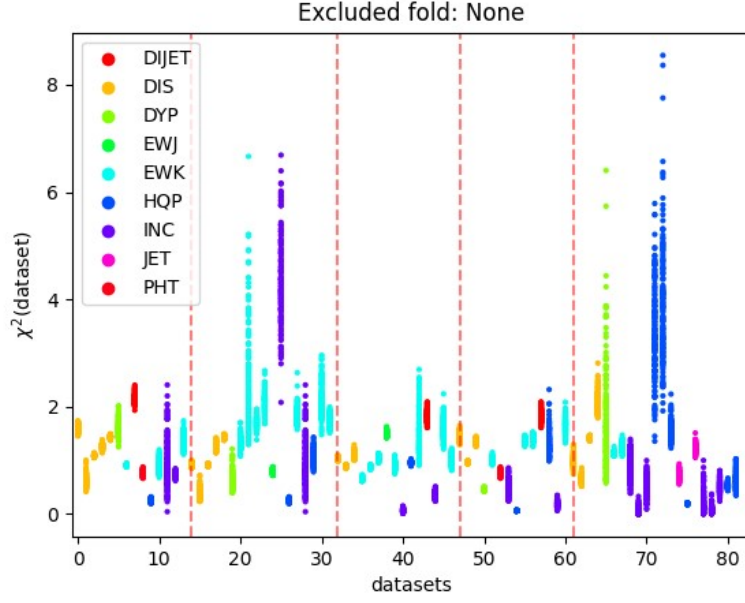

Figure 5.14: The plot shows the value of χ^2 computed on each dataset; the datasets are divided by the dotted line to highlight the four folds. This is a reference fit, where no dataset is excluded.

We repeat the plot in the four new cases and we show the results in Figure 5.15.

By comparison with Figure 5.14, we can see how the χ^2 on each dataset changes when that dataset is excluded.

In the first panel (where we exclude the fold 1), we can notice how, for a couple of datasets of the first fold, the χ^2 increases. The χ^2 also increases for a dataset in fold 2 (the one represented in violet, which reaches $\chi^2 = 8$ for some replicas), while it decreases for a couple of top quark datasets in the 'always included' group.

In the second panel, when the second fold it's excluded, we have an outstanding increasing of χ^2 for a dataset of the second fold: 'D0WMASY' (χ^2 goes from around 5 to around 20). The fact that this dataset contains only 10 data-points, together with the fact that fold 2 contains two out of three datasets regarding W asymmetry, could probably explain this increase of error; however, this also shows that this fold may not be well balanced.

In the third panel, there aren't significant differences with Figure 5.14, except for a slight increase of the χ^2 for some datasets of the third (excluded) fold.

Finally, in the forth panel, we have an increase of the χ^2 in a couple of datasets in the 'always included' group, while the excluded datasets of fold 4 seem to maintain their χ^2 value.

We will now focus on the fit obtained excluding fold 1 and fold 3, where the separation between best and worst χ^2 replicas is larger.
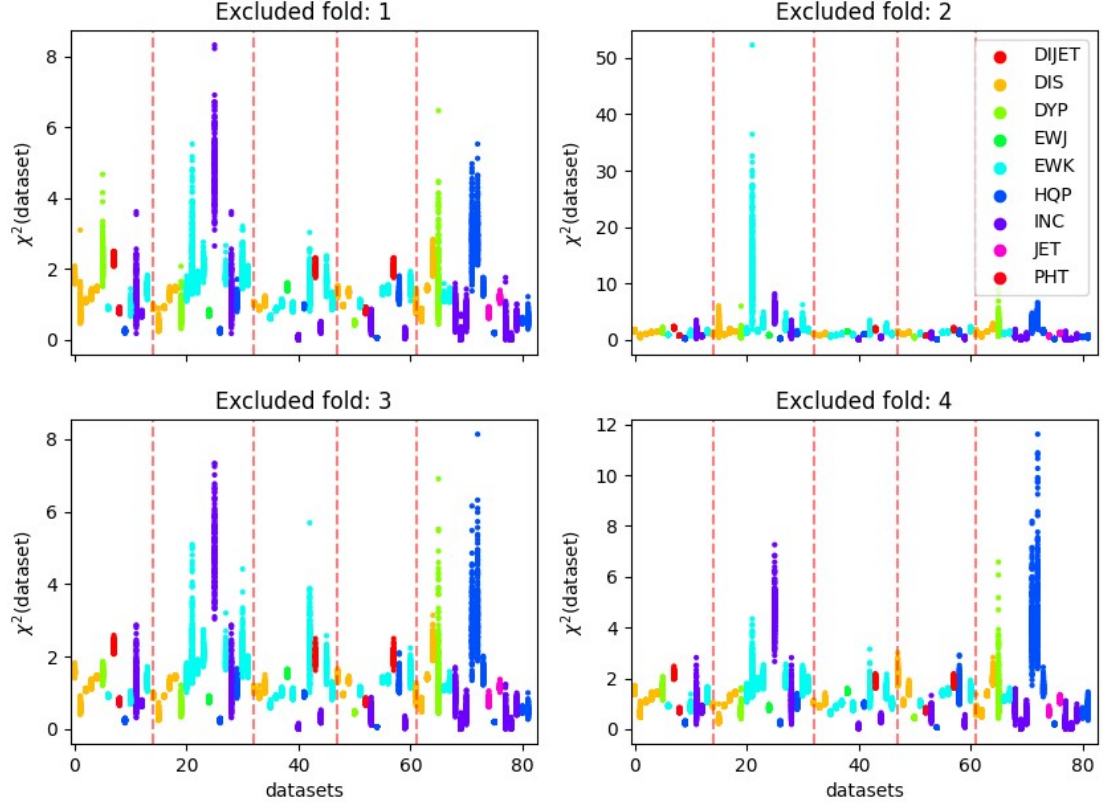
Figure 5.15: Same as in Figure 5.14 but for each of the four fit obtained excluding one fold at a time.

5.2.2 In-Depth Study of Fold 1 and Fold 3

In this section, we will focus on the fits done excluding respectively fold 1 and fold 3. As seen in Figure 5.5 and 5.7, in these fits we have a neater separation between best and worst χ^2 replicas, which we now want to investigate.

In order to understand the reason and the source of the separation, we compute the χ^2 per process, per experiment and per dataset, comparing the five replicas with best total χ^2 with the five replicas with worst total χ^2.

In Figure 5.16, the values of χ^2 for each experiment are presented, where the blue bars represent the χ^2 value of the five replicas with best total χ^2, while the red bars stand for the worst total χ^2 replicas. In both fits (the one obtained excluding fold 1 and the one excluding fold 3), for most of the experiments there's only a small χ^2 difference between best and worst replicas. The difference is bigger for the experiment D0, in both fits, while LHCb presents a more significant difference only when excluding fold 1.

In Figure 5.17, we present a plot similar to the previous one, only this time we compare χ^2 based on the process. Once again, we don't see big increases in the χ^2 of the worst replicas. For the fit done excluding fold 1, we see a bigger difference in χ^2 for the DY processes, while
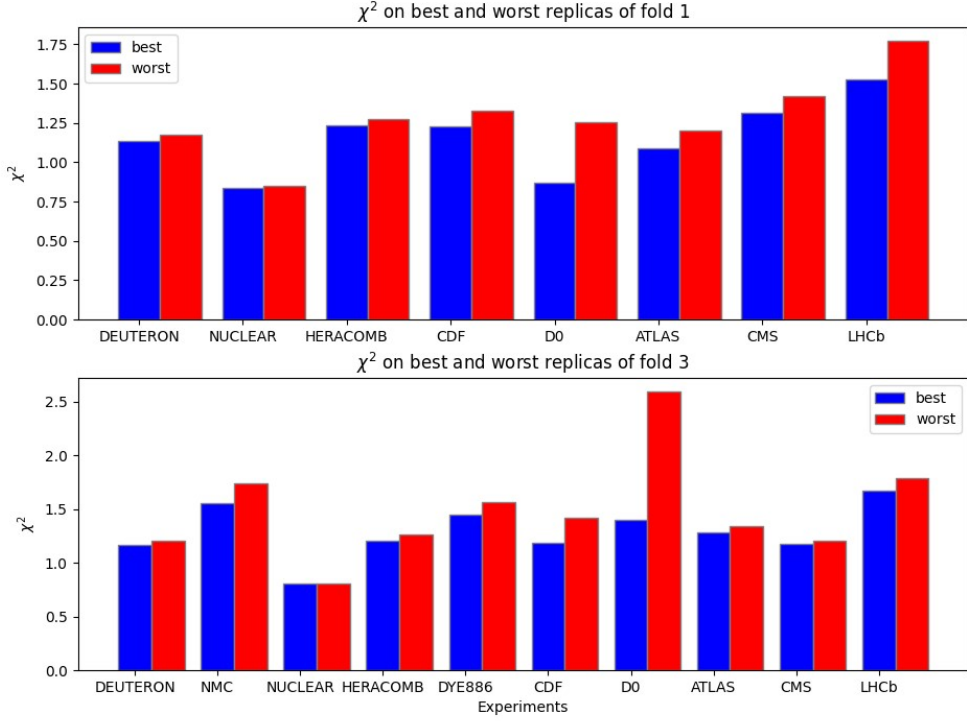
55

Figure 5.16: Comparison between χ^2 of each experiment between the best replicas (blue) and the worst replicas (red). On the upper panel, we show the results when excluding fold 1, while on the lower panel we exclude fold 3. The χ^2 in each case is computed only on the in-sample data-points.

this difference does not occur for fold 3. We can conclude that there is no particular process that determines the difference between best and worst replicas, instead, the reason of this difference may lay in the datasets. This means that the data-points which lead to outlier replicas do not come from a process in particular and we have to go on searching them in datasets.

We repeat the same analysis, this time looking at the χ^2 on the datasets. The results are presented in Figure 5.18, where we compare the χ^2 of the best and worst replicas for each of the datasets. The name of the datasets are not present in the plot, but they are divided in four folds by the dotted lines, the remaining datasets on the right are always included. In this plot, we see some significant difference in χ^2 value for certain datasets.

In order to identify which datasets are responsible for the difference in χ^2 between the best and the worst replicas, we plot the difference of the χ^2 on each dataset (this would be the difference between red and blue line of Figure 5.14). We show only the most significant datasets in Figure 5.19 and 5.20. In particular, we show the five datasets with highest and
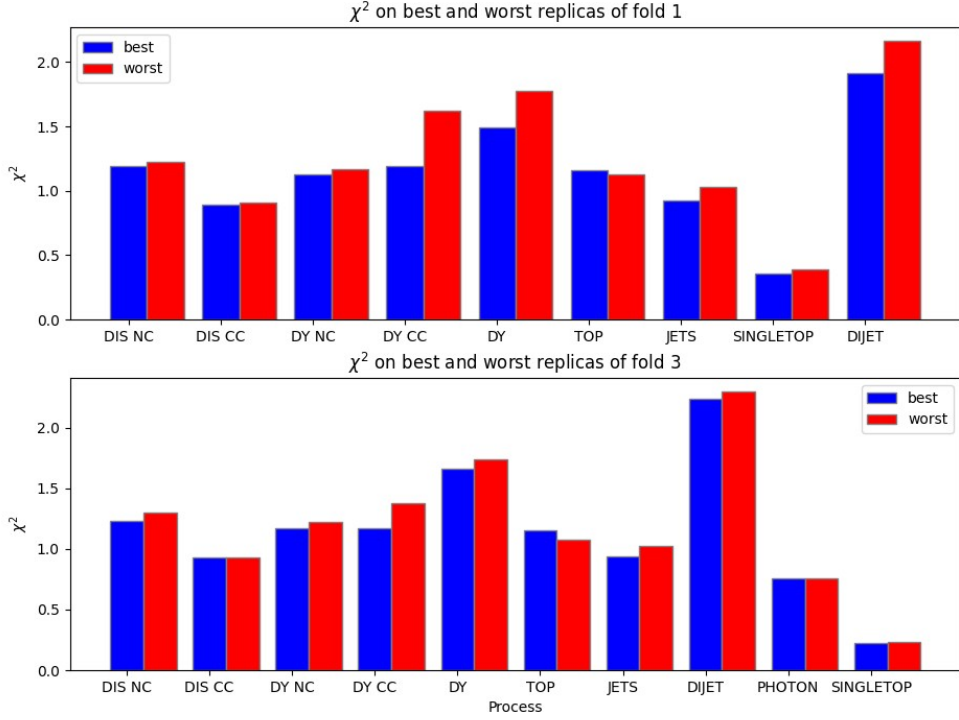
Figure 5.17: Comparison between χ^2 on the different processes between best (blue) and worst (red) replicas. On the upper panel are presented the results when excluding fold 1, while on the lower panel fold 3 is the excluded one. The χ^2 in each case is computed only on the data included in the fit.

lowest (negative) value of $\chi^2_{worst} - \chi^2_{best}$. The latter group contains datasets that actually are better interpolated by the worst χ^2 replicas than by the best ones. The datasets which present a positive difference, on the other hand, are the ones responsible for the difference in the total χ^2 of the replicas.

We can see that, in both the fits with excluded fold 1 and excluded fold 3, the dataset 'D0WMASY' is the one with the biggest difference $\chi^2_{worst} - \chi^2_{best}$. This is coherent with what we saw in Figure 5.16, where we noticed that experiments from D0 have the biggest effect in χ^2 shift. Furthermore, we have already noticed in Section 5.2.1 that this datasets has only ~ 10 data-points and it is one of the datasets with highest χ^2. So, it appears clear that this dataset is very sensitive to the difference between best and worst χ^2 replicas, and in particular it prefers the best χ^2 replicas, which have a higher peak and a lower tail. It is also clear that this dataset on its own has not enough data-points to be actively moving the fit. However, further analysis of this dataset and an understanding of why it favors the configuration of the best χ^2 replicas could be insightful.
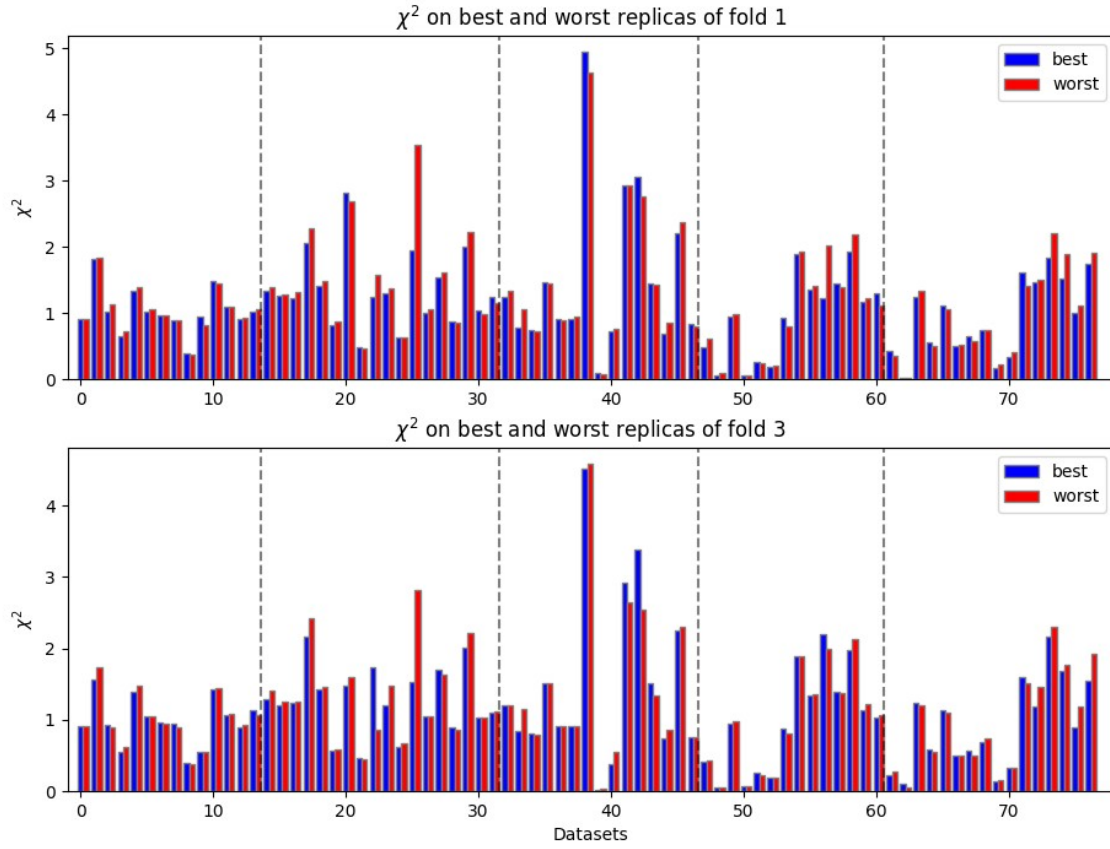
Figure 5.18: Comparison between χ^2 of each dataset on the best replicas (blue) and on the worst replicas (red). On the upper panel, we have the results when excluding fold 1, while on the lower panel we exclude fold 3. The dotted lines mimic the fold division, marking respectively the four folds and a fifth group of datasets which are always included in the fits.
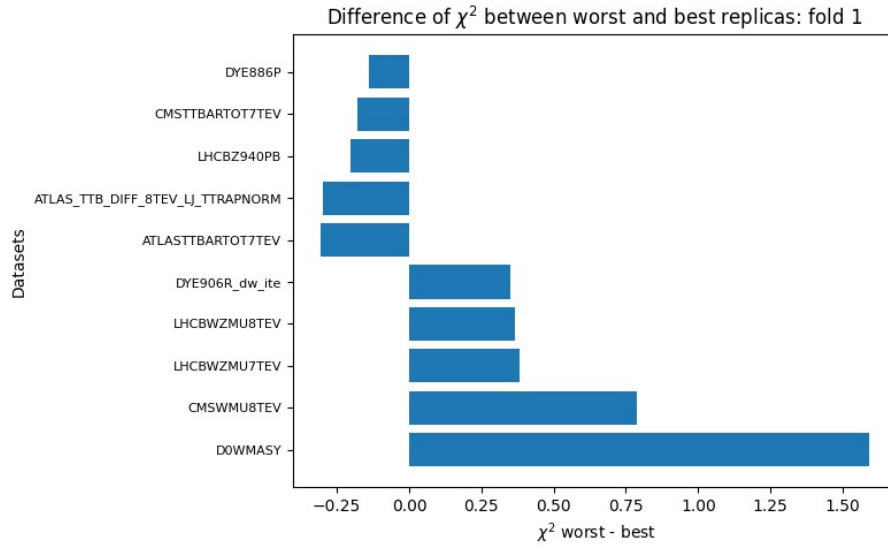
Figure 5.19: The plot contains the difference between χ^2 between worst and best replicas, computed for the most significant dataset, when excluding fold 1.
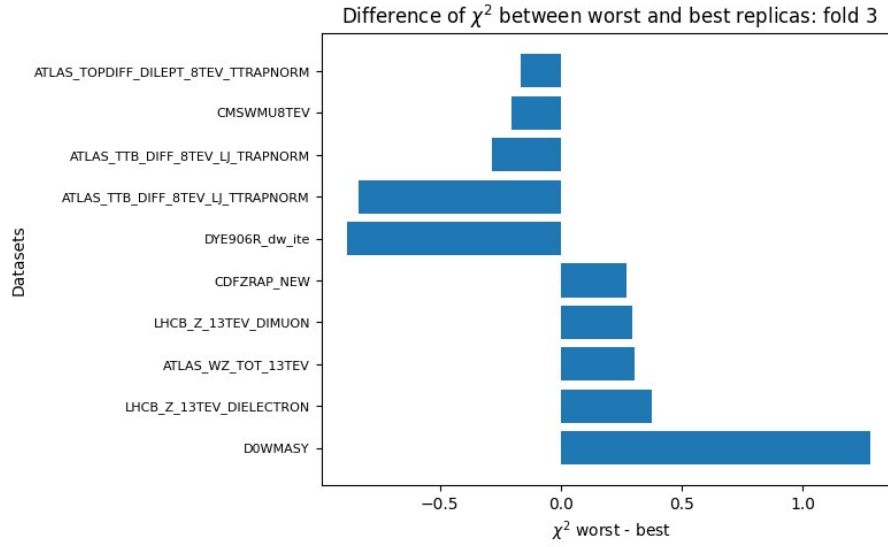

Figure 5.20: Same as for Figure 5.19, but excluding fold 3.

5.3 Impact of D0WMASY on the Fit

As said in the previous section, 'D0WMASY' appears to be the dataset with the larger difference in χ^2 between best and worst-χ^2 replicas. It is likely that some features in the shape of the best χ^2 replicas better represent this dataset. Looking at the Figures 5.5-5.8, we can identify these features in the height of the peak of the gluon, position on the x-axis of the peak and height of the tail. We will study the impact of the height of the peak on the χ^2 on 'D0WMASY'.

We plot, in Figure 5.21, the height of the gluon peak for each one of the five best and five worst-χ^2 replicas, along with the χ^2 of each replica on the dataset 'D0WMASY'.
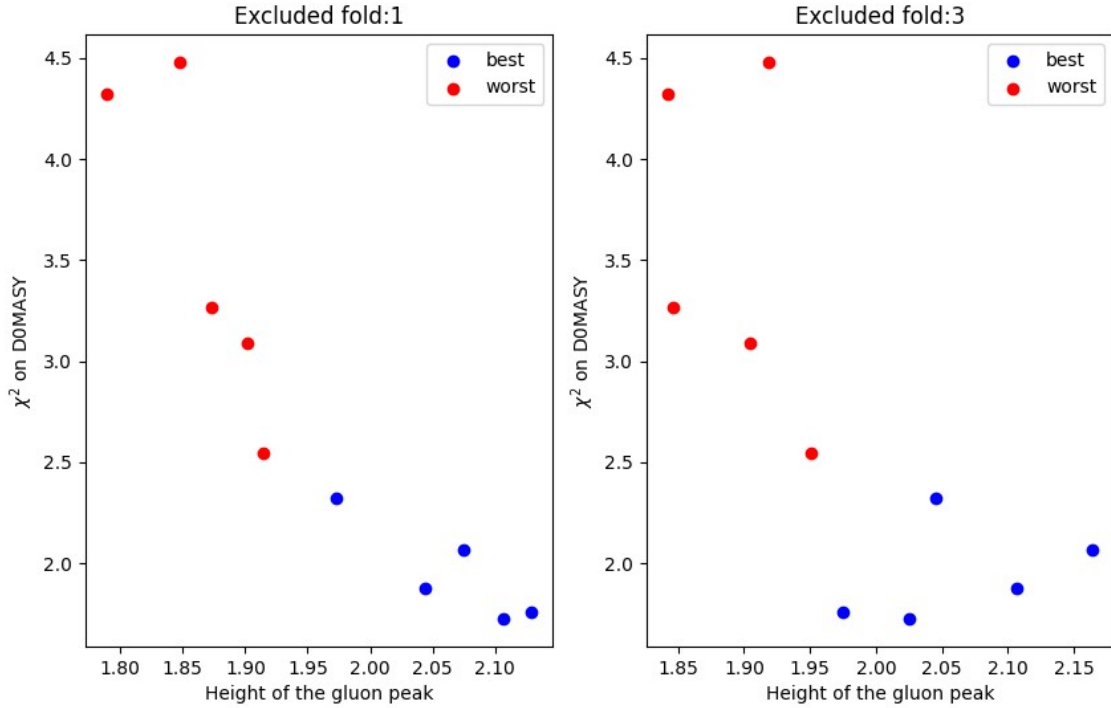
Figure 5.21: χ^2 on the dataset 'D0WMASY' in relation to the height of the gluon peak. We represent the five best-χ^2 replicas (in blue) and the five worst-χ^2 replicas (in red). On the left panel, we present the results of the fit obtained excluding fold 1, while on the right panel we exclude fold 3.

On the left panel, where we are excluding fold 1, we can hypothesise the presence of linear trend. On the right panel, on the other hand, there is a higher dispersion and it's less clear whether there is a linear trend.

To test whether there is a correlation between height of the peak of the gluon and χ^2 on 'D0WMASY', we will consider all the replicas and, to have a larger statistics, we consider

the fit 'NNPDF40_nnlo_as_01180_1000', which contains 1000 replicas. The results are presented in Figure 5.22, where each point represents a replica and the axis represents the χ^2 on D0WMASY and the height of the peak of the gluon. In this plot, the five best χ^2 replicas and five worst χ^2 replicas are marked by a star and colored in yellow and violet respectively. All the other points are also colored according to the total χ^2 of each replica.
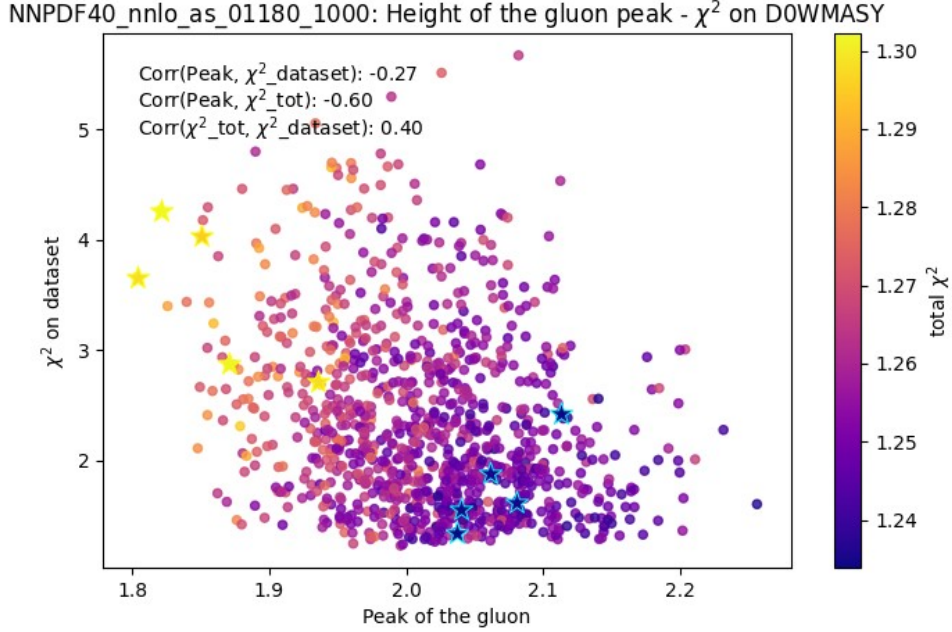

Figure 5.22: χ^2 on the dataset 'D0WMASY' in relation to the height of the gluon peak, for each of the 1000 replicas of the fit 'NNPDF40_nnlo_as_01180_1000'. The points are colored according to the total χ^2 value. The five replicas with highest and lowest value of χ^2 are marked by a star.

On the upper left, we show the correlations between the quantities in the plot.

The first value is the correlation between the height of the peak and the χ^2 on D0WMASY, which results to be corr$= -0.27$. This means that the previous plot, where only the five best and five worst χ^2 replicas were shown (Figure 5.21), was somehow magnifying this effect, which in actual fact is rather weak.

Let's also briefly comment the other correlation values in the plot. The second reported value is the correlation between the height of the peak and the total χ^2, so the correlation between the horizontal axis and the color scale. This value is -0.60. This means that replicas with a higher peak of the gluon have a lower χ^2. This was evident for the best and worst replicas, but now we have confirmed it for all the replicas and we have quantified the relation between height of the peak and goodness of the χ^2.

The last value in the plot quantify the correlation between the total χ^2 and the χ^2 on the

D0WMASY dataset. This would be the correlation between the vertical axis and the color scale and it results in a value of 0.40. This proves a non-negligible connection between these quantities, as we suspected from the big difference of the χ^2 on D0WMASY between best and worst χ^2 replicas in Figure 5.19 and 5.20, but it not a strong correlation.

We now want to see how the plot of Figure 5.1 changes in a fit without D0WMASY (which we will refer to as 'NNPDF40_nnlo_as_01180_1000_no_D0') and also how the correlations we took into consideration change.

We show the results of this new fit in Figure 5.23: in this case, the separation between best and worst χ^2 replicas seems to be smaller compared to the reference fit. However, this plot may be challenging to interpret at first glance because the differences compared to the reference fit (Figure 5.1) are subtle, as we are only excluding a small dataset. Furthermore, we need to remember that this new fit contains 100 replicas, which is significantly less than the reference fit, and a lower number of replicas usually decreases the separation effect between best and worst-χ^2 replicas.
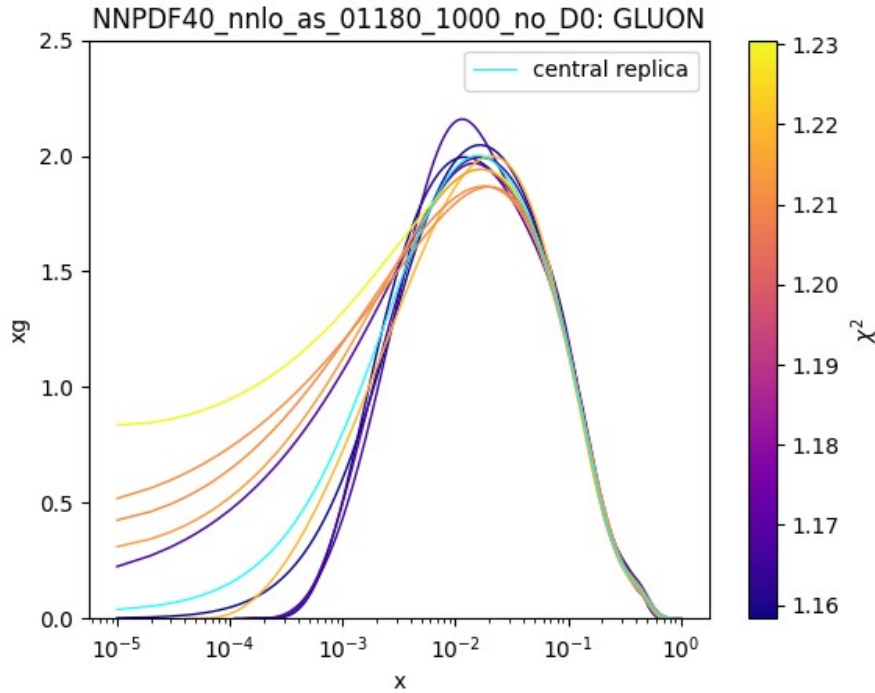
Figure 5.23: Same as Figure 5.1, but for a fit done excluding the dataset 'D0WMASY'.

Whether this decreased separation is a true effect or a consequence to the lower number of replicas has to be investigated. In order to do so, we can reproduce the scatterplot of Figure 5.22 for this new fit.

The results are shown in Figure 5.24. It is clear at first glance that the points in the plot are not correlated, in fact the correlation index between height of the peak and χ^2 on

D0WMASY is now only -0.08. Furthermore, the correlation between height of the peak and total χ^2 has lowered to -0.48 (it was -0.60 in the reference fit). This suggests that the decreased separation between best and worst χ^2 replicas of Figure 5.23 was a concrete effect of excluding D0WMASY.

Lastly, the correlation between the total χ^2 and the χ^2 on D0WMASY has decreased from 0.40 to 0.11, indicating that these two quantities are now essentially uncorrelated. This change may be due to D0WMASY being poorly fitted when it is excluded (we also noticed this in Section 5.2.1). We conclude that the fact that this correlation decreased does not disprove that this dataset plays a role in obtaining outlier replicas.
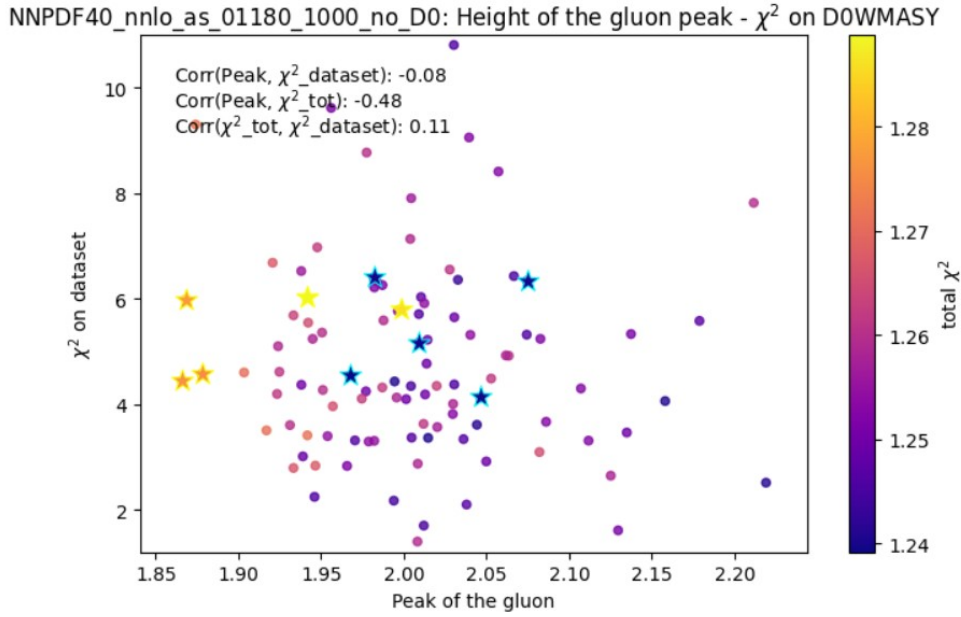
Figure 5.24: Same as in Figure 5.22, but for a fit done excluding D0WMASY.

5.3.1 Impact of Other Key Datasets

The plot of Figure 5.22 can be repeated for every dataset, reporting the correlation values in each case. When doing so, two datasets in particular stand out.

The first one is named 'HERACOMB_SIGMARED_B' and it is showed in Figure 5.25. As we can see from the Figure, in this case the correlation between χ^2 on the dataset and height of the gluon peak is very high (corr=-0.91). This means that this dataset has a substantial preference for a high peak gluon. Furthermore, this dataset seem to be more correlated with the total χ^2 compared to D0WMASY, in fact the correlation for this dataset is 0.54, against 0.40 for D0WMASY. One of the reasons can be that this dataset has more data-points (~ 27).

The second relevant dataset is called ATLAS_SINGLETOP_TCH_DIFF_8_TEV_T_RAP_-

Figure 5.25: χ^2 on the dataset 'HERACOMB_SIGMARED_B' in relation to the height of the gluon peak, for each of the 1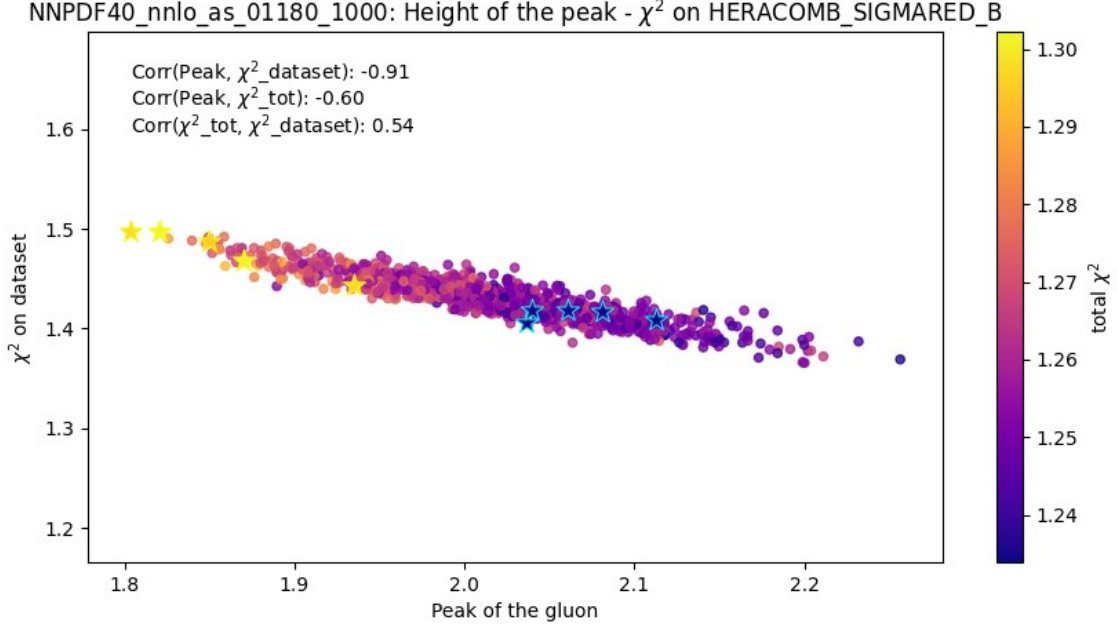000 replicas of the fit 'NNPDF40_nnlo_as_01180_1000'. The points are colored according to the total χ^2 value. The five replicas with highest and lowest value of χ^2 are marked by a star.

NORM and it is showed in Figure 5.26. In this case, the dataset χ^2 is positively correlated to the height of the gluon peak (corr= 0.63). This means that this dataset is pushing for a lower peak of the gluon. This effect has to be offset by other datasets (such as D0WMASY or HERACOMB_SIGMARED_B), since we know that the final χ^2 is negatively correlated to the height of the gluon peak (corr= -0.60 for the reference fit).

The correlation between total χ^2 and χ^2 on this dataset, coherently, is negative (corr= -0.32), this means that lower χ^2 values for this dataset are associated with higher χ^2 values for the rest of the data. This suggests potential tension between this dataset and others in the global fit.

It is important to consider that this dataset, which contains data from ATLAS for the single t process at $8\,TeV$, only has four data-points.

In both cases, it is interesting to notice that, although the correlation between height of the peak and χ^2 on the dataset is very strong, the difference in the dataset χ^2 value between the replicas with highest peak and the ones with lowest peak is not large (~ 0.1 in both cases). This subtle variation suggests that the neural network may not easily detect or prioritize this trend, as the change in χ^2 might be insufficient to influence the learning process in a significant way.

We now proceed, as in the previous section, producing two new fits: one excluding the dataset
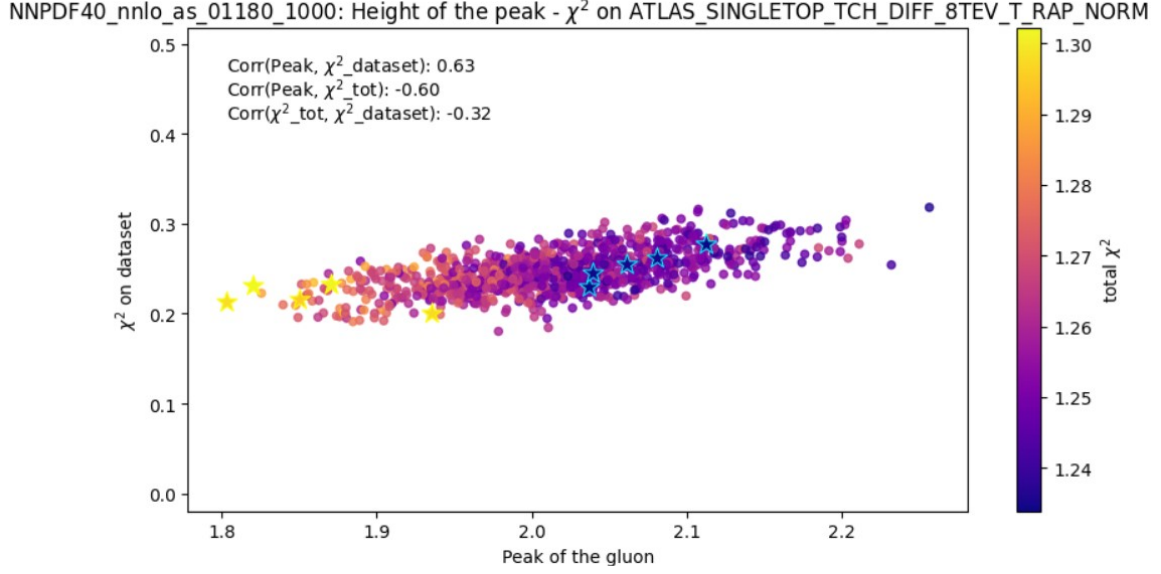
Figure 5.26: Same as in Figure 5.25, but for the dataset named 'ATLAS_SINGLETOP_TCH_-DIFF_8_TEV_T_RAP_NORM'

ATLAS_SINGLETOP_TCH_DIFF_8_TEV_T_RAP_NORM and one excluding HERACOMB_-SIGMARED_B.

We then want to replicate the plot of Figure 5.1 in these two cases. We expect that when excluding the ATLAS dataset, the separation between best and worst χ^2 replicas increases, since we are excluding a dataset that prefers a gluon with a lower peak. When excluding the HERACOMB dataset, on the other hand, we expect the separation to decrease, since we are excluding a dataset that is better represented by a higher peak of the gluon.

We show in Figure 5.27 the five best and five worst-χ^2 replicas in the fit obtained excluding ATLAS_SINGLETOP_TCH_DIFF_8_TEV_T_RAP_NORM, while in Figure 5.28 we present the same plot but for a fit done excluding HERACOM_SIGMARED_B.

The plots reveal a more pronounced separation in Figure 5.28, where the excluded dataset is HERACOMB_SIGMARED_B. Interestingly, this is contrary to our initial expectations. However, it is important to recognize that the differences between these fits might be subtle. This could be due to the fact that we are only excluding a small number of data points, and the relatively limited number of replicas may not capture the full effect.

To gain a deeper understanding of these trends, we can revisit the correlation indexes defined earlier. These values are presented in Table 5.3.

Examining the correlation between the height of the gluon peak and the total χ^2, we notice that although the visual separation between the best and worst replicas appeared larger for `HERACOMB_SIGMARED_B`, the actual correlation value is slightly lower in this case compared to `ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM` (-0.57 vs. -0.59). This supports our earlier observation.
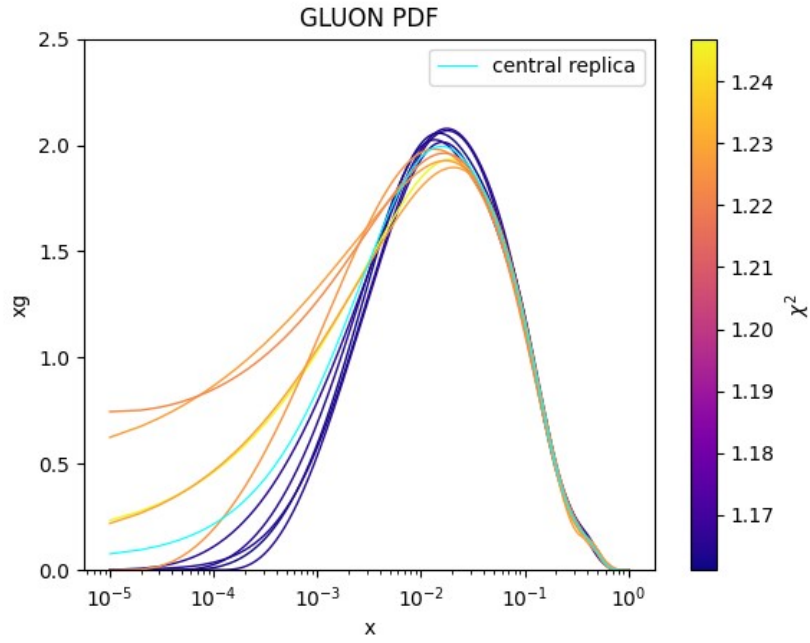
Figure 5.27: Same as for Figure 5.1, but for a fit containing 120 replicas and performed excluding ATLAS_SINGLETOP_TCH_DIFF_8_TEV_T_RAP_NORM.
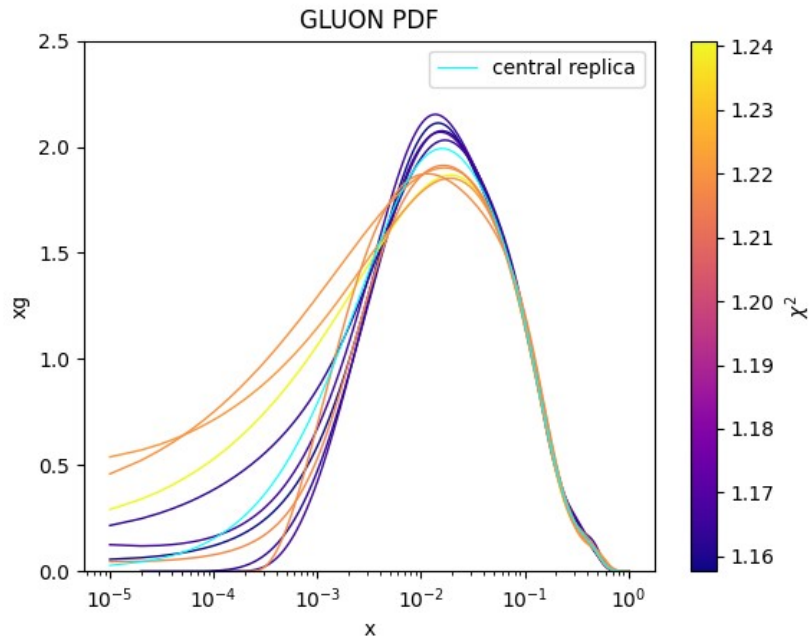

Figure 5.28: Same as for Figure 5.1, but for a fit containing 120 replicas and performed excluding HERACOM_SIGMARED_B.

Excluded Dataset	Corr(Peak, χ^2 Dataset)	Corr(Peak, χ^2)
`ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM`	0.59	-0.59
`HERACOMB_SIGMARED_B`	-0.93	-0.57

Table 5.3: Correlation values comparing fits excluding either `ATLAS_SINGLETOP_TCH_DIFF_-8TEV_T_RAP_NORM` or `HERACOMB_SIGMARED_B`. The first column shows the correlation between gluon peak height and χ^2 on the excluded dataset. The second column shows the correlation between peak height and total χ^2.

As for the correlation between the peak height and the χ^2 specific to the dataset, we see that this value remains almost unchanged after excluding `HERACOMB_SIGMARED_B` (-0.91 to -0.93). Similarly, excluding the ATLAS dataset does not significantly alter this correlation either, shifting only from 0.63 to 0.59.

Thus, we can conclude that while these two datasets do not influence the correlation between peak height and total χ^2 as strongly as D0WMASY, they maintain a stable correlation between their χ^2 and the peak height of the gluon, regardless of whether they are included or excluded in the fit. This means that the two datasets that we have discussed are one of the reason for outlier replicas. However, they may not be the sole responsible.

5.4 Defining Folds Based on Correlation with Gluon Peak Height

We now apply the method developed in previous sections to group datasets based on their correlation with the gluon peak height. Specifically, we calculate the correlation between the gluon peak height and the χ^2 for each dataset. The results show that many datasets exhibit little to no correlation with the gluon height, while some show a significant positive or negative correlation.

To identify relevant correlations, we set a threshold: correlations greater than 0.4 or less than -0.4 are considered significant. Datasets with correlations above 0.4 form the group labeled *positive_corr*, while those with correlations below -0.4 form the group *negative_corr*. Tables 5.4 and 5.4 presents the datasets in each group along with their respective correlation values. Despite these two groups contain a similar number of datasets, the number of data-points they include is significantly different: the group *negative_corr*, in fact, contains substantially more data-points.

We will now perform two new fits: one excluding the datasets in *positive_corr* and another excluding those in *negative_corr*. These fits will be named after the group of datasets excluded.

5.4.1 Comparison between Fits: *positive_corr* vs *negative_corr*

Figures 5.29 and 5.30 show the five best and five worst χ^2 replicas, respectively, for the fits excluding datasets in the *positive_corr* and *negative_corr* groups. From the plots, we observe
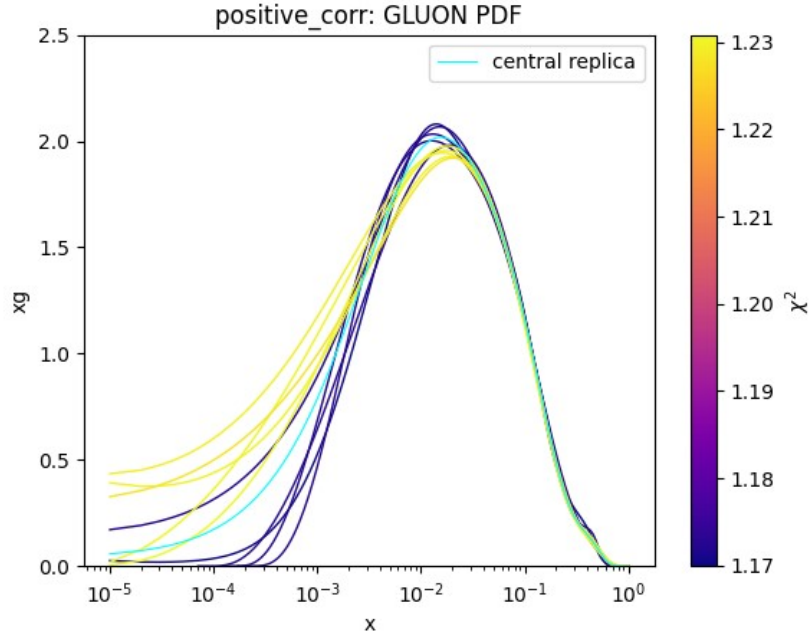
Figure 5.29: Same as in Figure 5.1, but for a fit containing 90 replicas and performed excluding the datasets in *positive_corr*.
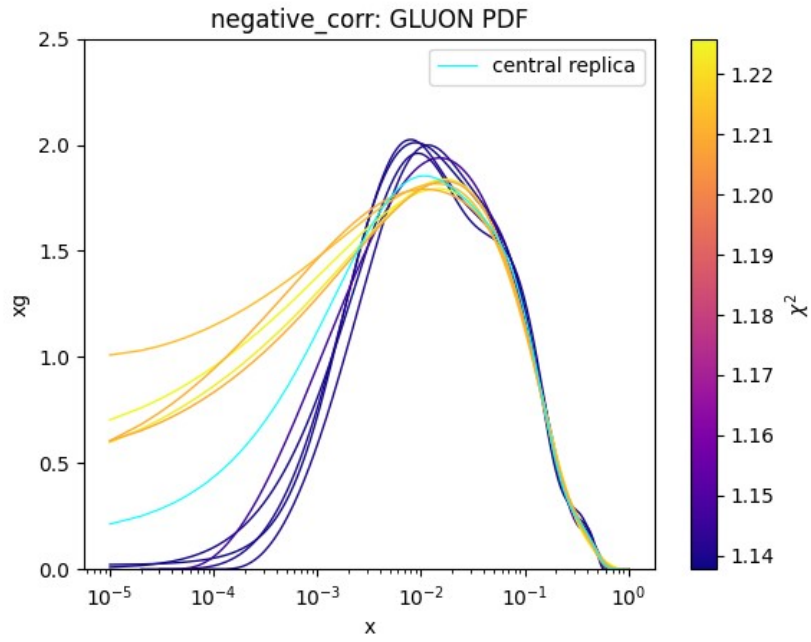

Figure 5.30: Same as in Figure 5.1, but for a fit containing 90 replicas and performed excluding the datasets in *negative_corr*.

Dataset	Correlation Peak
ATLASZPT8TEVMDIST	0.45
ATLASPHT15_SF	0.44
ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM	0.63
ATLASLOMASSDY11EXT	0.61
ATLAS_SINGLETOP_TCH_DIFF_7TEV_TBAR_RAP_NORM	0.43
CMSZDIFF12	0.62

Table 5.4: Positive correlation datasets and the correlation between χ^2 on each dataset and height of the gluon peak.

Dataset	Correlation Peak
NMC	-0.56
HERACOMB_SIGMARED_B	-0.91
HERACOMBNCEP575	-0.43
LHCB_Z_13TEV_DIELECTRON	-0.47
ATLAS_SINGLETOP_TCH_DIFF_7TEV_T_RAP_NORM	-0.47
ATLAS_SINGLETOP_TCH_DIFF_8TEV_TBAR_RAP_NORM	-0.53
CMS_1JET_8TEV	-0.50

Table 5.5: Negative correlation datasets and the correlation between χ^2 on each dataset and height of the gluon peak.

a greater separation in the *negative_corr* fit, while the replicas in *positive_corr* are much less dispersed. This might seem counterintuitive, as one would expect the separation to decrease when excluding datasets that drive a higher peak. However, it's important to consider that *negative_corr* contains a substantial amount of data. When this data is excluded, the overall quality of the fit diminishes, leading to increased dispersion.

Furthermore, the fact that the height of the peak in Figure 5.30 is lower than in Figure 5.29 is reassuring and aligns with expectations. To reinforce this point, we can also examine the histogram of the gluon peak height for the two fits, in Figure 5.31. The entire distribution for the fit excluding *positive_corr* is shifted toward higher peak values. In the plots, the mean peak height and the height of the central replica are marked. In both cases, the central replica has a lower peak than the mean of the peak distribution.

We can also look at the correlation between total χ^2 and height of the peak, which, as we know from the previous sections, gives an important quantitative information about the fit. This correlation value is -0.61 for the fit done excluding *positive_fit*, while it's -0.54 for the fit where we exclude *negative_fit*. We recall that this value is -0.60 when we include all the datasets. These values are in line with what we would expect: when excluding the datasets that prefer a lower gluon, the anti-correlation between total χ^2 and height of the gluon becomes stronger, while excluding the datasets that prefer a higher peak, the anti-correlation decrease.

Although the plots of Figure 5.29 and 5.30 seemed to contradict the initial hypothesis, the

more reliable and objective correlation values confirm it.

We now want to focus on the χ^2 distribution of the two fits we are analyzing. As we pointed out, from Figure 5.29 and 5.30, it is noticeable that the χ^2 is lower when we exclude the datasets in *negative_corr*. This suggest that the latter datasets actually contain the data-points that generate outlier replicas.

In Figure 5.32, we show the distribution of χ^2 for the fit done excluding *negative_corr* (left panel) and *positive_corr* (right panel). For the first, the χ^2 ranges from 1.14 to 1.22, while the second has χ^2 ranging from 1.17 to 1.23. In the plots it is also showed the mean of the χ^2 distribution and the χ^2 value of the central replica. Both these values are lower for the fit named *negative_corr*.

We can further investigate the source of the χ^2 difference between the two fits looking at the χ^2 values of each dataset. In particular, we compute the difference between χ^2 in the fit *positive_corr* and χ^2 in the fit *negative_corr*, for each dataset.

We show in Figure 5.33 the ten most significant datasets. Among these, the ones displaying a negative value of the difference are in tension with the datasets of *positive_corr*. Analogously, the ones displaying a positive value of the difference, which are mainly dataset from top pair production processes, are in tension with the datasets in *negative_corr*. This means that datasets from top production, which do not contain many data-points, seem to prefer a gluon with a lower peak.

We see that the first two datasets with lowest difference (NMC, CMS_1JET_8TEV)), are in the group *negative_corr*. The third is D0WMASY. This dataset, which we previously investigated, was found to have no significant correlation with the height of the peak. However, the difference in its χ^2 value indicates an overall preference for a higher gluon peak.

5.4.2   The effect of the NMC Dataset

NMC is one of the datasets in *negative_corr* and, as we have seen, it is significantly more in agreement with the results of the fit *positive_corr* compared to the results of *negative_corr*.

For this dataset, which contains ~ 300 data-points, we can qualitative observe how its data-points drive the fit.

In fact, we notice a particular feature which is present whenever we exclude the dataset 'NMC'. If we look at the fits where this dataset is not in-sample (namely, the fit done excluding fold 1 and the fit done excluding *negative_corr*, respectively Figure 5.5 and 5.30), they both present replicas with a particular feature: a rapid change of slope at $x \sim 0.5$. This is interestingly compatible with the x range of the dataset NMC.

This means that this dataset is particularly important for obtaining a smooth gluon PDF around $x \sim 0.5$. When it is excluded, certain replicas can assume a non physical shape. This gives a lower χ^2 on included data but higher on excluded ones (as it happens for fold 1, Figures 5.5 and 5.9), which indicates overfitting.
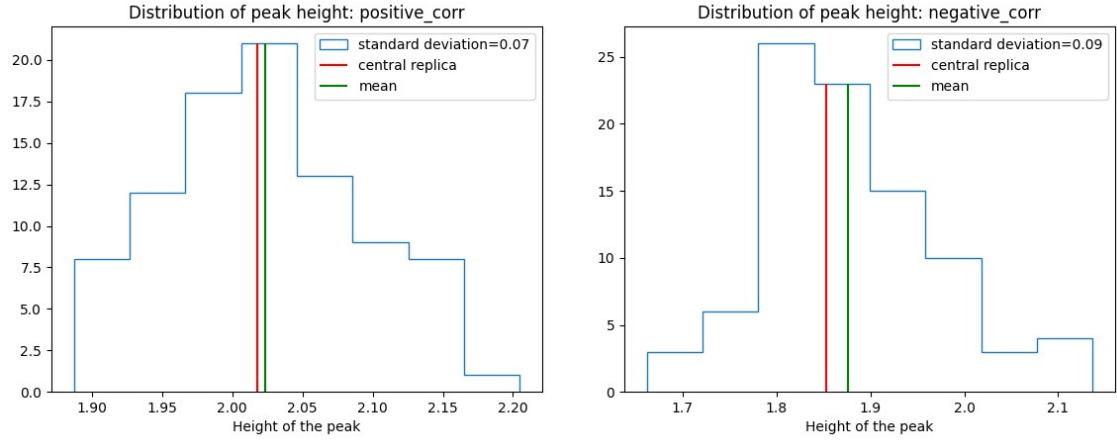
Figure 5.31: The histogram shows the distribution of height of the gluon peak for the replicas of the fit *positive_corr* (left panel) and *negative_corr* (right panel). In red, it's marked the height of the peak of the central replica. In green, the mean height of the peak. The plot also include the standard deviation value of the distribution.
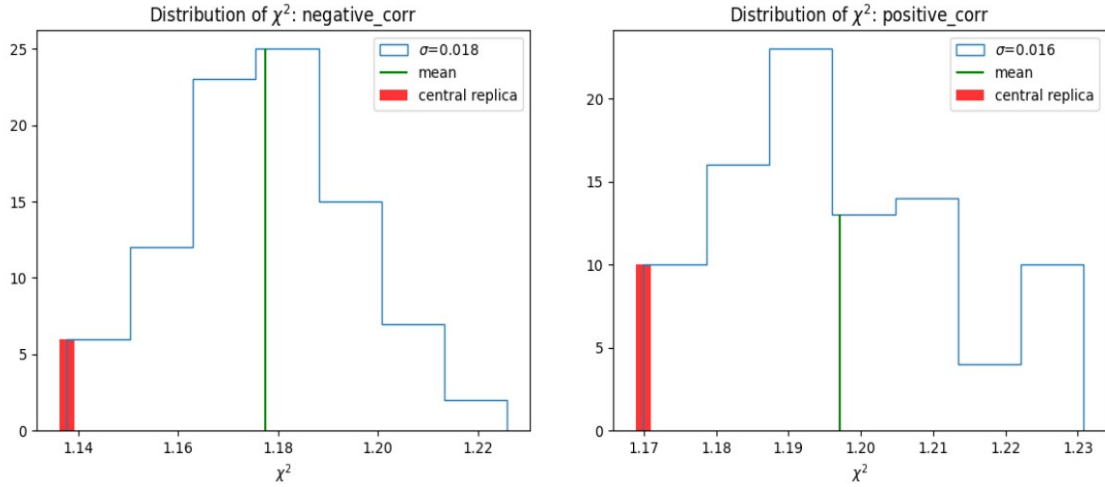
Figure 5.32: χ^2 distribution for the fit done excluding the datasets in *negative_corr* (left panel) and excluding the ones in *positive_corr* (right panel). It is also reported the mean and standard deviation of the distribution, as well as the χ^2 of the central replica.

Figure 5.33: Difference between the χ^2 in the fit *positive_corr* and in the fit *negative_corr*, for each dataset.

5.5 Interpretation in Retrospect of K-Folding Results

Now that we have identified which datasets have an impact on the height of the gluon peak and we have determined a numerical value to quantify the separation between best and worst χ^2 replicas, we can reanalyze the results of the fits presented in section 5.2.1.
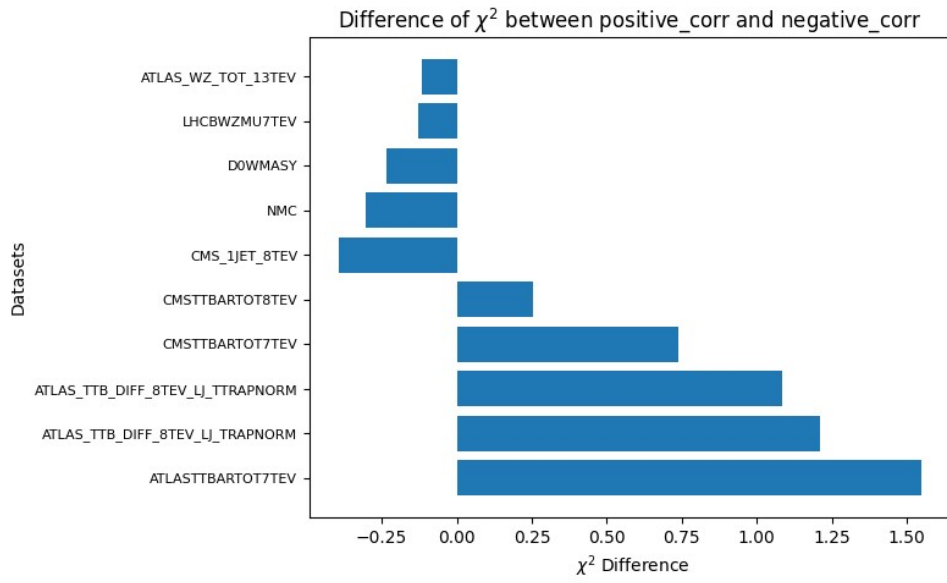
First of all, we highlight which of the datasets in each fold are also contained in either *positive_corr* (colored in green) or *negative_corr* (in red).

Fold 1	Fold 2
ATLAS_2JET_7TEV_R06	ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM
ATLAS_SINGLETOP_TCH_DIFF_8TEV_T_RAP_NORM	ATLAS_WP_JET_8TEV_PT
ATLASPHT15_SF	ATLASTTBARTOT7TEV
ATLASZPT8TEVMDIST	ATLASWZRAP11CC
CMS_SINGLETOP_TCH_TOT_7TEV	ATLASZHIGHMASS49FB
CMSTTBARTOT7TEV	CDFZRAP_NEW
CMSWEASY840PB	CMS_TTBAR_2D_DIFF_MTT_TRAP_NORM
DYE886P	CMSTTBARTOT7TEV
HERACOMB_SIGMARED_B	CMSWMASY47FB
HERACOMBCCEM	D0WMASY
HERACOMBNCEP460	DYE886R_dw_ite
LHCBZEE2FB_40	HERACOMBCCEP
NMC	HERACOMBNCEP575
NTVNBDMNFe_dw_ite	LHCB_Z_13TEV_DIELECTRON
	LHCBWZMU7TEV
	NMCPD_dw_ite
	NTVNUDMNFe_dw_ite
	HERACOMB_SIGMARED_B

Table 5.6: Datasets for Folds 1 and 2. Green cells represent positive correlation datasets, and red cells represent negative correlation datasets.

Fold 3	Fold 4
ATLAS_SINGLETOP_TCH_DIFF_7TEV_T_RAP_NORM	ATLAS_SINGLETOP_TCH_DIFF_7TEV_TBAR_RAP_NORM
ATLAS_SINGLETOP_TCH_R_13TEV	ATLAS_SINGLETOP_TCH_R_7TEV
ATLAS_WM_JET_8TEV_PT	ATLASPHT15_SF
ATLASLOMASSDY11EXT	ATLASWZRAP36PB
ATLASWZRAP11CF	BCDMSP_dwsh
ATLASZPT8TEVYDIST	CHORUSNUPb_dw_ite
BCDMSD_dw_ite	CMS_2JET_7TEV
CHORUSNBPb_dw_ite	CMS_SINGLETOP_TCH_R_8TEV
CMS_2JET_7TEV	CMSDY2D11
CMS_SINGLETOP_TCH_R_13TEV	CMSTOPDIFF8TEVTTRAPNORM
CMSWMU8TEV	CMSZDIFF12
D0ZRAP_40	DYE605_dw_ite
HERACOMBNCEP820	HERACOMBNCEP920
LHCB_Z_13TEV_DIMUON	LHCBZ940PB
LHCBWZMU8TEV	

Table 5.7: Datasets in Fold 3 and Fold 4. Datasets with positive correlation are highlighted in green, and datasets with negative correlation are highlighted in red.

Then, we report the correlation between total χ^2 and height of the peak for each of the fits, the results are in Table 5.8. The values in the table confirm our initial remarks regarding these fits: when excluding fold 1 and fold 3 we encounter a higher separation, while the separation decreases excluding fold 2 or fold 4.

The correlation value of -0.61, obtained excluding fold 1, can be explained noticing that fold

Excluded Fold	Correlation (Chi2, Peak)
Fold 1	-0.61
Fold 2	-0.46
Fold 3	-0.51
Fold 4	-0.47

Table 5.8: Correlation between χ^2 and gluon peak height for each fit.

1 contains both datasets with positive and negative correlation, which may counterbalance, resulting in a correlation value very similar to the one of the reference fit (-0.60).

The value of -0.46, obtained excluding the second fold, can be explained by the presence of three datasets with negative correlation in this fold, which determine a weakening of the anti-correlation effect, similar to the situation showed in the previous section.

Fold 3 contains one dataset of *positive_corr* and one of *negative_corr*, making it rather balanced. The correlation value in the table is -0.51, suggesting that the exclusion of the dataset with negative correlation has a bigger impact than the exclusion of the *positive_corr* dataset.

Lastly, in fold 4 we exclude, among others, three folds with positive correlation. Unexpectedly, the correlation in the table results to be -0.47, when it was expected to be stronger. It is possible that, in this situation, effects from other datasets become significant.

Chapter 6

Conclusions

In this thesis, we addressed the challenge of introducing explainable AI into the NNPDF framework.

In particular, we investigated the reasons behind the behavior of the replicas of the gluon PDF. As Figure 5.1 shows, the replicas with the best agreement with the data (which have a higher peak and lower tail), are distinctly separated from the ones with the worst agreement with the data (which have a lower peak and higher tail), and the central replica is not the one with the best agreement with the data.

We measured the agreement to the data with the *experimental* χ^2. Considering all the replicas, we calculated that the correlation between the height of the peak and χ^2 is -0.60.

To explain the fact that the lowest-χ^2 replicas are outliers, we first hypothesized that they were *overfitting* the data, if this was true, they would fit data fluctuations and not generalize well to excluded data. However, this was proven to be false. In fact, we performed four different fits excluding different datasets each time, and we found no evidence of an increased χ^2 on the excluded datasets for the replicas with the best-χ^2. The four groups of datasets used in this step are called *folds*, and they are the same ones used for the determination of the model's hyperparameters, which determine the particular fitting methodology adopted. Performing an in-depth study on the folds, we found that fold 1 and fold 3 presented more clearly the effect we wanted to investigate, and thus we focused on those.

Looking at the fit done excluding fold 1 and the fit done excluding fold 3, we checked whether the best replicas were fitting some processes particularly well compared to others, and we concluded this is not the case. We repeated this check on experiments and found that the best replicas fitted particularly well data from the D0 experiment. Comparing datasets, we found that the dataset 'D0WMASY' was significantly more in agreement with the best replicas compared to the worst replicas.

We hypothesized that this dataset could be the reason for the higher-peak best replicas. To prove this, we calculated the correlation between the χ^2 of each replica on D0WMASY and the height of its peak. This showed a weak correlation between the two values, equal to

75

−0.27. While looking only at the best and worst replicas suggested that this dataset preferred a higher peak, the larger picture obtained considering all the replicas disproved this idea.

We then focused on the dataset with the highest and lowest correlation between χ^2 on the dataset and the height of the peak. These datasets are ATLAS_SINGLETOP_TCH_DIFF_-8TEV_T_RAP_NORM and HERACOMB_SIGMARED_B, and their correlations are 0.63 and −0.91, respectively. We produced two new fits, excluding one of the two datasets in each fit, and computed the correlation between total χ^2 and height of the peak. Keeping in mind that this value was −0.60 for the reference fit, we obtained −0.59 excluding the dataset positively correlated and −0.57 excluding the dataset with a negative correlation. This is aligned with what one would expect: excluding a dataset that is negatively correlated (which means that it has a preference for a higher peak of the gluon), the correlation between total χ^2 and height of the peak decreases. However, the difference between the two fits is minimal, likely because only a few data points distinguish one from the other.

To obtain a stronger difference, we defined two groups of datasets: *positive_corr* and *negative_corr*. The first contains all the datasets with a correlation between χ^2 on the dataset and height of the peak higher than 0.40, while the second contains all the datasets with a correlation value lower than −0.40. It is important to note that 'negative_corr' contains significantly more data points. We then performed two fits excluding one of these groups at a time (and named the fit after the excluded group), and we showed the best and worst replicas in each of the two cases (see Figure 5.30, 5.29).

The plots suggest that the correlation between total χ^2 and height of the peak could be higher for the fit 'negative_corr', where we have excluded the datasets that prefer a higher peak. If true, this would be the opposite of what we expect. However, the correlation value between total χ^2 and height of the peak disproves this qualitative consideration and consolidates the formulated hypothesis. In fact, the correlation value turns out to be −0.61 when excluding positive_corr and −0.54 when excluding negative_corr.

We conclude that the datasets in the group negative_corr are the reason for the outlier replicas. Since this group contains more data points than the group positive_corr, the replicas with a higher peak have an overall lower χ^2, while the replicas with a lower peak have a worse value of χ^2 because they accommodate the preference of fewer data points. The presence of datasets that prefer a lower peak also explains why the central replica does not coincide with the one with the best χ^2.

Bibliography

[1] Alexandre Deur, Stanley J. Brodsky, and Guy F. de Teramond. The QCD Running Coupling. *Nucl. Phys.*, 90:1, 2016.

[2] Robert Hofstadter and Robert W. McAllister. Electron scattering from the proton. *Phys. Rev.*, 98:217–218, Apr 1955.

[3] E. D. Bloom, D. H. Coward, H. DeStaebler, J. Drees, G. Miller, L. W. Mo, R. E. Taylor, M. Breidenbach, J. I. Friedman, G. C. Hartmann, and H. W. Kendall. High-energy inelastic $e - p$ scattering at 6° and 10°. *Phys. Rev. Lett.*, 23:930–934, Oct 1969.

[4] Jun Gao, Lucian Harland-Lang, and Juan Rojo. The Structure of the Proton in the LHC Precision Era. *Phys. Rept.*, 742:1–121, 2018.

[5] Michelangelo L. Mangano. Introduction to QCD. In *1998 European School of High-Energy Physics*, pages 53–97, 1998.

[6] P. Nason. Introduction to perturbative QCD. In *11th Jorge Andre Swieca Summer School on Particle and Fields*, pages 409–486, 1 2001.

[7] Yuri L. Dokshitzer, Valery A. Khoze, Alfred H. Mueller, and S. I. Troian. *Basics of perturbative QCD*. 1991.

[8] P. Nason. Introduction to QCD. *Conf. Proc. C*, 9705251:94–149, 1997.

[9] Morad Aaboud et al. Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. *JHEP*, 09:020, 2017.

[10] A. Aktas et al. Measurement and QCD analysis of the diffractive deep-inelastic scattering cross-section at HERA. *Eur. Phys. J. C*, 48:715–748, 2006.

[11] P. Bagnaia et al. Evidence for $Z^0 \to e^+ e^-$ at the CERN $\bar{p}p$ Collider. *Phys. Lett. B*, 129:130–140, 1983.

[12] G. Arnison et al. Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV. *Phys. Lett. B*, 122:103–116, 1983.

[13] J. Pumplin, D. Stump, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung. Uncertainties of predictions from parton distribution functions. 2. The Hessian method. *Phys. Rev. D*, 65:014013, 2001.

[14] Stefano Forte and Stefano Carrazza. Parton distribution functions. 8 2020.

[15] Stefano Forte, Lluis Garrido, Jose I. Latorre, and Andrea Piccione. Neural network parametrization of deep inelastic structure functions. *JHEP*, 05:062, 2002.

[16] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, José I. Latorre, Andrea Piccione, Juan Rojo, and Maria Ubiali. A determination of parton distributions with faithful uncertainty estimation. *Nuclear Physics B*, 809(1):1–63, 2009.

[17] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[18] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[19] J Gao, L Harland-Lang, and J Rojo. The structure of the proton in the lhc precision era. *International Journal of Modern Physics A*, 32(5):1730014, 2017.

[20] J Pumplin, D Stump, J Huston, et al. New generation of parton distributions with uncertainties from global qcd analysis. *Journal of High Energy Physics*, 2002(07):012, 2002.

[21] Stefano Forte, Lluis Garrido, Jose I Latorre, and Andrea Piccione. Neural network parametrization of deep-inelastic structure functions. *Journal of High Energy Physics*, 2002(05):062, 2002.

[22] Stefano Forte, Giovanni Ridolfi, and Juan Rojo. Parton distribution functions. *Progress in Particle and Nuclear Physics*, 112:103771, 2020.

[23] R. G. Roberts. *The Structure of the Proton: Deep Inelastic Scattering.* Cambridge Monographs on Mathematical Physics. Cambridge University Press, 1990.

[24] Richard D. Ball et al. The path to proton structure at 1% accuracy. *Eur. Phys. J. C*, 82(5):428, 2022.

[25] Alessandro Candido, Stefano Forte, and Felix Hekhorn. Can $\overline{\text{MS}}$ parton distributions be negative? *JHEP*, 11:129, 2020.

[26] E. E. Chambers and R. Hofstadter. Structure of the proton. *Phys. Rev.*, 103:1454–1463, Sep 1956.

[27] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015.

[28] Richard D. Ball, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, José I. Latorre, Andrea Piccione, Juan Rojo, and Maria Ubiali. Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering. *Nuclear Physics B*, 823(1–2):195–233, December 2009.

[29] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Christopher S. Deans, Luigi Del Debbio, Stefano Forte, Alberto Guffanti, Nathan P. Hartland, José I. Latorre, Juan Rojo, and Maria Ubiali. Parton distributions for the lhc run ii. *Journal of High Energy Physics*, 2015(4), April 2015.

[30] George F. Sterman and Steven Weinberg. Jets from Quantum Chromodynamics. *Phys. Rev. Lett.*, 39:1436, 1977.

[31] Yuri L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber. Better jet clustering algorithms. *JHEP*, 08:001, 1997.

[32] Aurore Courtoy, Joey Huston, Pavel Nadolsky, Keping Xie, Mengshi Yan, and C.-P. Yuan. Parton distributions need representative sampling. *Physical Review D*, 107(3), February 2023.

[33] Richard D. Ball, Juan Cruz-Martinez, Luigi Del Debbio, Stefano Forte, Zahari Kassabov, Emanuele R. Nocera, Juan Rojo, Roy Stegeman, and Maria Ubiali. Response to "Parton distributions need representative sampling". 11 2022.

[34] Richard D. Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Patrick Groth-Merrild, Alberto Guffanti, Nathan P. Hartland, Zahari Kassabov, José I. Latorre, Emanuele R. Nocera, Juan Rojo, Luca Rottoli, Emma Slade, and Maria Ubiali. Parton distributions from high-precision collider data: Nnpdf collaboration. *The European Physical Journal C*, 77(10), October 2017.

[35] M. Arneodo. Nuclear effects in structure functions. *Phys. Rept.*, 240:301–393, 1994.

[36] D. F. Geesaman, T. Saito, and A. W. Thomas. The Nuclear EMC effect. *Ann. Rev. Nucl. Part. Sci.*, 45:337–390, 1995.

[37] A. Bodek and B. G. Ritchie. Electron-deuteron and electron-proton scattering: The effects of fermi motion, binding, and nucleon off-shellness. *Phys. Rev. D*, 23:1070–1080, 1981.

[38] K. J. Eskola, H. Paukkunen, and C. A. Salgado. Eps09: A new generation of nlo and lo nuclear parton distribution functions. *Journal of High Energy Physics*, 2009(4), 2009.

[39] R. Keith Ellis, W. James Stirling, and B. R. Webber. *QCD and collider physics*, volume 8. Cambridge University Press, 2 2011.

Ringraziamenti

Ringrazio il mio relatore Prof. Stefano Forte per avermi introdotta a questo progetto stimolante ed avermi dato importanti consigli e suggerimenti. Lo ringrazio per la sua disponibilità e guida durante tutto il lavoro di tesi.

Ringrazio il mio correlatore Dr. Andrea Barontini che con pazienza mi ha introdotta all'uso del codice. Il suo aiuto è stato fondamentale per portare a termine questo progetto.

Ringrazio anche la mia famiglia, i miei amici e il mio ragazzo per il sostegno che mi hanno dato in questi anni.