



Artificial neural network modelling of generalised parton distributions

H. Dutrieux^{1,a}, O. Grocholski^{2,b}, H. Moutarde^{1,c}, P. Sznajder^{3,d}

¹ IRFU, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

² Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

³ National Centre for Nuclear Research (NCBJ), Pasteura 7, 02-093 Warsaw, Poland

Received: 22 December 2021 / Accepted: 13 March 2022 / Published online: 23 March 2022
© The Author(s) 2022, corrected publication 2022

Abstract We discuss the use of machine learning techniques in effectively nonparametric modelling of generalised parton distributions (GPDs) in view of their future extraction from experimental data. Current parameterisations of GPDs suffer from model dependency that lessens their impact on phenomenology and brings unknown systematics to the estimation of quantities like Mellin moments. The new strategy presented in this study allows to describe GPDs in a way fulfilling theory-driven constraints, keeping model dependency to a minimum. Getting a better grip on the control of systematic effects, our work will help the GPD phenomenology to achieve its maturity in the precision era commenced by the new generation of experiments.

1 Introduction

Generalised parton distributions (GPDs) [1–5] are widely recognised as one of the key objects to explore the structure of hadrons. They encompass information coming from one-dimensional parton distribution functions (PDFs) and elastic form factors (EFFs). GPDs allow for a hadron tomography [6–8], where densities of partons carrying a fraction of hadron momentum are studied in the plane perpendicular to the hadron’s direction of motion. GPDs also provide access to the matrix elements of the energy-momentum tensor [2, 3, 9], making it possible to evaluate the total angular momentum and “mechanical” properties of hadrons, like pressure and shear stress at a given point of space [10–13].

Experimental access to GPDs is mainly possible thanks to exclusive processes occurring on hadrons remaining coher-

ent after a hard scale interaction. Some notable processes of this type are deeply virtual Compton scattering (DVCS) [3], time-like Compton scattering (TCS) [14] and deeply virtual meson production (DVMP) [4, 15]. All of them allow to study transitions of hadrons from one state to another, with a unique insight into changes taking place at the partonic level. GPDs have been primarily studied in lepton production experiments, in particular those conducted in JLab, DESY and CERN, and are key objects of interest in programmes of future electron-ion colliders, like EIC [16, 17], EicC [18] and LHeC [19]. Many data sets have already been collected and reviewed for DVCS [20, 21] and DVMP [22], while the first measurement for TCS has been completed recently [23].

Although several datasets are already available for fits, the extraction of GPDs is far from being satisfactory. The main reasons are:

- (i) sparsity of available information. GPDs are multidimensional functions, so one needs much more data to constrain them in the phase-space of kinematic variables, comparing to e.g. one-dimensional PDFs. Furthermore, one needs to cover this phase-space by data collected for various processes and experimental setups, which is required to distinguish between many types of GPDs and contributions coming from various quark flavours and gluons.
- (ii) complexity of extraction. In order to fully benefit from available sources of information about GPDs, such as data collected in exclusive measurements, PDFs and EFFs for boundary conditions, lattice QCD, one needs to know and implement links connecting those sources with GPDs. The extraction of GPDs requires a deconvolution of amplitudes measured in exclusive processes, which is not trivial and in some cases does not even possess a unique solution [24]. Both the evolution of GPDs and the description of exclusive processes must

^a e-mail: herve.dutrieux@cea.fr

^b e-mail: oskar.grocholski@desy.de

^c e-mail: herve.moutarde@cea.fr

^d e-mail: pawel.sznajder@ncbj.gov.pl (corresponding author)

be understood to a level allowing for a robust extraction. This requires a substantial effort and a careful design of tools aggregating GPD-related theory developments.

- (iii) model dependency. Currently available phenomenological GPD models, like GK [25–27] and VGG [28–31], use similar Ansätze with a rigid form and therefore cannot be considered as diverse sources for the estimation of model uncertainty. The severity of such a model dependency and its impact on e.g. the extraction of orbital angular momentum have never been studied in a systematic way, partially because there were no tools to do so.

The answer of the GPD community to those problems so far has been as follows. More data sensitive to GPDs will be delivered by a next generation of experiments. These experiments, performed in many laboratories and with multiple setups, will provide the much needed input to the GPD phenomenology. We witness a substantial progress in the description of exclusive processes and in the systematic use of GPD evolution equations [32]. The development of lattice QCD techniques is accelerating, as proved by Ref. [33]. The PARTONS [34] and GeParD [35] projects provide aggregation points for GPD-related developments and practical know-how in phenomenology methods. Techniques to stress model dependency are developed for the extraction of amplitudes, the latter used to access the “mechanical” properties of nucleons [36,37]. However, to the best of our knowledge, no substantial progress has been recently made to address the problem of model dependency of GPDs extractions.

In this Article, we discuss new ways of modelling GPDs which provide flexible parameterisations inspired by artificial neural network (ANN) techniques. These parameterisations may be used as tools to study the model dependency affecting extractions of GPDs and derived quantities, like orbital angular momentum of hadrons. We address aspects of neural network optimisation, such as over-learning and convergence. Our work aims in providing a better grip on the control of systematic effects, much needed in front of the precision era of GPD extractions.

We note that methods providing effectively nonparametric modelling are already widely used with a great success to determine PDFs, see e.g. Refs. [38–41]. It is therefore natural to adopt them and the associated extraction strategies in the analysis of GPDs. The challenges preventing their direct application are mainly due to the fact that GPDs depend on more kinematic variables in comparison to PDFs, and require to enforce extra nontrivial theoretical constraints. In this Article we show for the first time how to meet these challenges.

The Article is organised as follows. In Sect. 2 we provide the theoretical background for our line of research and summarize the theory driven constraints that GPD parameterizations must obey. The modelling of GPDs in (x, ξ) -space

is given in Sect. 3, while that in (β, α) -space is discussed in Sect. 4. We provide the summary in Sect. 5.

2 Theoretical background

GPDs are functions of three variables: the average longitudinal momentum fraction carried by the active parton, x , the skewness ξ , which describes the longitudinal momentum transfer, and the four-momentum transfer to the hadron target, t . GPDs also depend on the factorisation scale, μ_F , being the variable entering evolution equations. A knowledge of these equations allows to run the evolution starting from a reference scale, $\mu_{F,0}$, where GPD models are defined, to any μ_F . We therefore only consider GPDs at $\mu_{F,0}$, and in the following we suppress the μ_F dependence for brevity. For illustration we will only consider the GPD $H(x, \xi, t)$ for a quark of unspecified flavour, but we note that the discussion can be easily extended for other GPDs. $H(x, \xi, t)$ is a real and ξ -even function as a consequence of QCD invariance under discrete symmetries. Without loss of generality, we only consider positive ξ in this work.

GPD models must fulfil a set of theory driven constraints:

- (i) reduction to the PDF $q(x)$ in the forward limit:

$$H(x, \xi = 0, t = 0) \equiv q(x), \quad (1)$$

- (ii) polynomiality [42–44], which is required to keep GPDs invariant under Lorentz transformation. The property states that any Mellin moment of a GPD, \mathcal{A}_n , is a fixed-order polynomial in ξ :

$$\begin{aligned} \mathcal{A}_n(\xi, t) &= \int_{-1}^1 dx x^n H(x, \xi, t) \\ &= \sum_{\substack{j=0 \\ \text{even}}}^n \xi^j A_{n,j}(t) + \text{mod}(n, 2) \xi^{n+1} A_{n,n+1}(t), \end{aligned} \quad (2)$$

where we call $A_{n,j}(t)$ a Mellin coefficient. We note that $\mathcal{A}_0(\xi, t) = A_{0,0}(t) \equiv F_1(t)$ is the Dirac EFF form factor. The polynomiality “entangles” the x and ξ dependencies.

- (iii) positivity constraints [45–53], which are inequalities ensuring positive norms in the Hilbert space of states. Because of the variety of these inequalities, we do not quote them all here. Instead, we refer the Reader to the aforementioned references, and only note that the constraints typically involve several types of GPDs. In the following we will illustrate how to deal with the positivity constraints with this simple inequality [45, 46],

$$|H(x, \xi, t)| \leq \sqrt{q\left(\frac{x+\xi}{1+\xi}\right)q\left(\frac{x-\xi}{1-\xi}\right)\frac{1}{1-\xi^2}}, \tag{3}$$

which is applicable in the DGLAP region, i.e. for $x > \xi$.

For the sake of completeness, we write a direct relation between Mellin moments, which are extensively discussed in this Article, and conformal moments given by:

$$C_n(\xi, t) = \xi^n \int_{-1}^1 dx C_n^{3/2}\left(\frac{x}{\xi}\right) H(x, \xi, t), \tag{4}$$

where $C_n^{3/2}(x)$ are Gegenbauer polynomials:

$$C_n^{3/2}(x) = \sum_{\substack{k=0 \\ \text{even}}}^n c_{n,k} x^{n-k}, \tag{5}$$

with coefficients

$$c_{n,k} = (-1)^{k/2} \frac{\Gamma(n - k/2 + 3/2)}{\Gamma(3/2)(k/2)!(n - k)!} 2^{n-k}. \tag{6}$$

A finite number of Mellin moments is needed to express a single conformal moment of a fixed order,

$$C_n(\xi, t) = \sum_{\substack{k=0 \\ \text{even}}}^n c_{n,k} \xi^k \mathcal{A}_{n-k}(\xi, t), \tag{7}$$

and *vice versa*, a finite number of conformal moments is needed to express a single Mellin moment,

$$c_{0,0} \mathcal{A}_0(\xi, t) = C_0(\xi, t), \tag{8}$$

$$c_{n,0} \mathcal{A}_n(\xi, t) = C_n(\xi, t) - \sum_{\substack{k=2 \\ \text{even}}}^n c_{n,k} \xi^k \mathcal{A}_{n-k}(\xi, t). \tag{9}$$

The conformal moments conveniently diagonalise the evolution equations at leading-order (LO) [54], but they also appear in the modelling of GPDs, like for instance in Ref. [55].

GPDs may be equivalently [56] defined in the double distribution representation of β and α variables, which is related to (x, ξ) -space by the Radon transform:

$$f(x, \xi) = \int d\Omega F(\beta, \alpha), \tag{10}$$

where $d\Omega = d\beta d\alpha \delta(x - \beta - \alpha\xi)$ and $|\alpha| + |\beta| \leq 1$. Working in (β, α) -space allows us to relatively easily fulfil both the reduction to PDFs and the polynomiality property. However, any projection of experimental observables using GPD models defined in that space, or *vice versa*, any attempt to find free parameters of such models from experimental data, requires either the Radon transform or its inverse counterpart. This is typically done numerically, which severely

slows down computations. In some cases the Radon transform can be performed analytically, but it requires a rather simple form of double distributions and may be numerically unstable due to delicate cancellations of small numbers, typically occurring at small ξ (see e.g. the $\mathcal{O}(\xi^{-5})$ factor in Eq. (27) of Ref. [57]). The inverse Radon transform poses a challenge of its own, as reported for instance in Ref. [58].

We close this section with the remark that the core of this work is related to preserving proper forward limits, polynomiality and positivity of GPDs models. This primarily does not involve the t variable, which in general is not relevant in our study and as in other analyses of this type will be suppressed for brevity. In summary, we are left with the problem of modeling GPDs in 2D spaces: either (x, ξ) or (β, α) .

3 Modelling in (x, ξ) -space

3.1 Principles of modelling

The polynomiality property suggests that the $A_{n,j}$ coefficients introduced in Eq. (2) provide a convenient basis to describe GPDs in a flexible modelling based on multiple parameters. The coefficients are not expected to be correlated, making them true degrees of freedom for GPDs. Another reason to choose this basis is the relation between Mellin and conformal moments indicated in Sect. 2. This relation allows one to perform the LO evolution of GPDs in a fast and straightforward way. Finally, the strategy advocated here is a convenient way to incorporate calculations of Mellin moments performed by lattice QCD in the modelling of GPDs.

Because there are infinitely many $A_{n,j}$ coefficients, for a practical use we only consider a subset of them. This is justified by the fact that the evaluation of higher order Mellin moments becomes increasingly sensitive to the numerical noise: if $H(x, \xi)$ is a bounded summable function, its Mellin moment $\mathcal{A}_n(\xi)$ tends to zero when n gets to infinity, in which case it starts to drown in the noise in any evaluation. This prevents the extraction of high order Mellin moments from experimental data without adding prior assumptions. Evaluation on the lattice also suffers from increased uncertainty at higher orders. Therefore, even if in principle the full series of Mellin moments is required to unambiguously reconstruct a function, this situation almost never appears in practical situations. The general case to consider is actually when only a finite number of Mellin moments are known. An extra regularization is thus necessary to restore the uniqueness of the function defined by its Mellin moments (in the present case, this extra regularization is introduced by the choice of a polynomial basis or an ANN basis as described below). By linearity, we point out that this argument still holds true when dealing with conformal moments instead of Mellin moments.

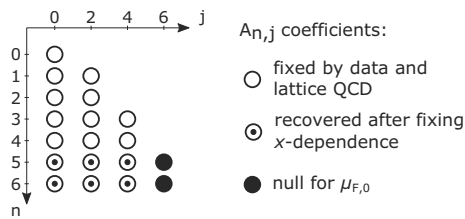


Fig. 1 The sketch indicates three groups of the polynomiality expansion coefficients, $A_{n,j}$, appearing in the modelling after fixing the truncation parameter to $N = 4$. The coefficients that we assume to be constrained by data and lattice QCD are denoted by \circ , those that are null, but only at the initial factorisation scale are denoted by \bullet , while those calculable from \circ after fixing the recovery of x -dependence are denoted by \odot

In our modelling, we only let free the coefficients $A_{n,j}$ appearing in Mellin moments up to and including the order $n \leq N$. The value of N is arbitrary and it controls the flexibility of models, just like the order of polynomials or the size of artificial neural networks used in modern parameterisations of PDFs. In other words, N fixes the number of degrees of freedom. For even values of N , which for brevity we only consider in this study, this number is $(1 + N/2)(2 + N/2) - 1$.

The coefficients $A_{n,j}$ for $n > N$ and $j \leq N$ are recovered from those for $n \leq N$ by a specific reconstruction procedure of the x -dependence that is discussed in the following. On the contrary, we assume that the coefficients $A_{n,j}$ vanish for $n > N$ and $j > N$. Note that this is true at $\mu_{F,0}$ where we are defining the model. For $\mu_F \neq \mu_{F,0}$, evolution will make these once null coefficients to have non-zero values, but only for $j \leq n + \text{mod}(n, 2)$ since evolution preserves the polynomiality property. We summarise this paragraph with a sketch shown in Fig. 1.

The reconstruction of the x -dependence from a number of $A_{n,j}$ coefficients is the *moment problem* known in mathematics. Let us address it by expressing GPDs in a way suggested by the polynomiality property:

$$H(x, \xi) = \sum_{\substack{j=0 \\ \text{even}}}^N f_j(x) \xi^j. \tag{11}$$

Here, $f_j(x)$ is a function of x that satisfies the following relation to Mellin coefficients:

$$\int_{-1}^1 dx x^n f_j(x) = A_{n,j}, \tag{12}$$

where $A_{n,j} = 0$ for $j > n + \text{mod}(n, 2)$. Since $H(x = \pm 1, \xi) = 0$ and considering the polynomial form of Eq. (11), $f_j(x)$ has to vanish at the end points:

$$f_j(-1) = f_j(1) = 0. \tag{13}$$

Equations (12) and (13) are essential to constrain $f_j(x)$ and therefore to obtain valid GPD models. In the following we

will try two bases for $f_j(x)$, i.e. two forms of this function: one based on monomials, which has already been studied in the literature (see Ref. [56] and references therein), and one inspired by the ANN technique. A given basis fixes the bridge between the picture of GPDs given by Mellin moments (ξ -dependent integrals over x) and the picture where GPDs are explicitly given as functions of x and ξ . We note that the selection of a given basis and the necessity of keeping a finite N introduces some model-dependency. We however expect this bias to become arbitrarily small for sufficiently large N .

It is important to note that GPDs may exhibit a singularity at $x = \xi = 0$, which corresponds to the singularity of PDF at $x = 0$. However, this singularity does not affect the evaluation of Mellin moments. Since here we are modeling GPDs with a polynomial in ξ , see Eq. (11), we are not able to effectively build a model that exhibits the singularity at only $x = \xi = 0$ (and not at $x = 0, \xi \neq 0$). It is a general flaw of this type of modelling, which motivated us to find ANN-inspired GPDs in (β, α) -space, which we present in Sect. 4. Still, a direct modelling of (x, ξ) -space can be used for a subset of non-singular GPD models, like the pion model [59,60] used for a demonstration in Sect. 3.4. In addition, one can consider a reference GPD model, $H_0(x, \xi)$, with a correct $H_0(x, \xi)/q(x)$ behaviour for $\xi \rightarrow 0$, and constrain a polynomial model, $H_1(x, \xi)$, such that $H_0(x, \xi) + H_1(x, \xi)$ correctly reproduces $\xi > 0$ data and e.g. lattice results for Mellin moments. This kind of hybrid modelling may be plausible for phenomenology, as models defined in (x, ξ) -space can be directly compared to data, and therefore their usage can severely speed up minimization procedures.

3.2 Polynomial basis

According to the Stone-Weierstrass theorem, any continuous function defined on a compact set can be uniformly approximated by a polynomial to any desired degree of precision. Applying this theorem to the two-dimensional function $H(x, \xi)$ suggests to express $f_j(x)$ that is defined in the $|x| \leq 1$ interval by

$$f_j(x) = \sum_{i=0}^{N+2} w_{i,j} x^i. \tag{14}$$

Here, $w_{i,j}$ are coefficients multiplying the $x^i \xi^j$ monomials in the global expression of $H(x, \xi)$. The order of polynomial (14) is $N + 2$, and it is the minimal order required to satisfy both Eqs. (12) and (13). We note that the order may be higher, which provides extra degrees of freedom available in the modelling.

Mellin coefficients can be evaluated from monomial ones in the following way:

$$\forall_{\substack{\text{even } n \\ n}} A_{n,j} = \sum_{\substack{i=0 \\ \text{even}}}^{N+2} w_{i,j} \int_{-1}^1 dx x^{i+n} = \sum_{\substack{i=0 \\ \text{even}}}^{N+2} w_{i,j} \frac{2}{1+i+n}, \tag{15}$$

$$\forall_{\substack{\text{odd } n \\ n}} A_{n,j} = \sum_{\substack{i=1 \\ \text{odd}}}^{N+1} w_{i,j} \int_{-1}^1 dx x^{i+n} = \sum_{\substack{i=1 \\ \text{odd}}}^{N+1} w_{i,j} \frac{2}{1+i+n}, \tag{16}$$

where we remind that N is even, and where Eq. (13) yields:

$$\sum_{i=0}^{N+2} w_{i,j} = \sum_{i=0}^{N+2} (-1)^i w_{i,j} = 0. \tag{17}$$

To find the opposite relations we need to solve Eqs. (15), (16) and (17) for $w_{i,j}$. We express all these equations in the matrix form:

$$\mathbf{A}_j = \mathcal{C} \cdot \mathbf{w}_j. \tag{18}$$

Here, $\mathbf{A}_j = (A_{0,j}, \dots, A_{n,j}, \dots, A_{N,j}, 0, 0)^T$ and $\mathbf{w}_j = (w_{0,j}, \dots, w_{i,j}, \dots, w_{N+2,j})^T$, and the length of each vector is $N + 3$. The matrix \mathcal{C} is:

$$\mathcal{C} = \begin{pmatrix} c_{0,0} & \dots & c_{0,i} & \dots & c_{0,N+2} \\ \dots & \dots & \dots & \dots & \dots \\ c_{n,0} & \dots & c_{n,i} & \dots & c_{n,N+2} \\ \dots & \dots & \dots & \dots & \dots \\ c_{N+2,0} & \dots & c_{N+2,i} & \dots & c_{N+2,N+2} \end{pmatrix}, \tag{19}$$

where:

$$c_{n,i} = \begin{cases} \frac{2}{1+i+n} & n \leq N \text{ and even } n+i, \\ 0 & n \leq N \text{ and odd } n+i, \\ (-1)^{in} & \text{otherwise.} \end{cases} \tag{20}$$

This matrix is only characterized by the parameter N . Therefore, as soon as this parameter is fixed the matrix is ready for the inversion (for instance numerically), and this can be done only once in the modelling procedure.

We note that the presented realisation of GPDs in the monomial basis of minimal order $N + 2$ naturally arises in the context of the so-called dual parameterisation [10] after its initial formal representation has been recast as a convergent series. This parameterisation is physically motivated by an infinite series of t -channel exchanges. We also note that any attempt of modelling GPDs in the (x, ξ) -space based on classical orthogonal polynomials of x , like Gegenbauer polynomials, will lead to the same conclusions as presented in this section. This is because any monomial of x can be expressed by a finite number of classical orthogonal polynomials.

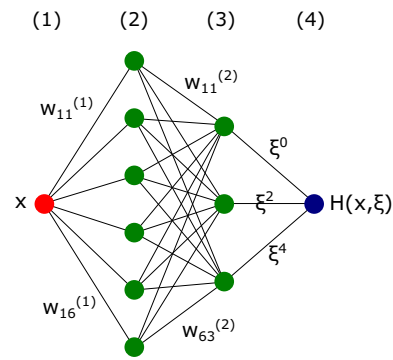


Fig. 2 Scheme of an exemplary artificial neural network used to represent a single GPD. The example is for the truncation parameter $N = 4$

3.3 Artificial neural network basis

In the case of ANNs, universal approximation theorems (see e.g. Ref. [61]) for arbitrary width or depth of the neural network ensure that a large enough network is able to accurately represent any continuous function on a compact set. The graphical representation of an exemplary network used to construct valid GPD models is shown in Fig. 2.

The exemplary network is made out of four layers: input and output layers ((1) and (4), respectively), and two hidden layers ((2) and (3)). The signal is processed through the network as follows:

- (i) The neuron in the input layer receives the value of x and distributes it to all neurons in the first hidden layer (second layer of the network: (2)) via a set of connections. A single weight is associated to each of those connections that is denoted by $w_{i,j}^{(1)}$, where in our example $i = 1, \dots, 6$.
- (ii) Each neuron in the hidden and output layers processes the incoming signal according to this equation:

$$o_k^{(l)} = \varphi_k^{(l)} \left(b_k^{(l)} + \sum_i o_i^{(l-1)} w_{i,k}^{(l-1)} \right), \tag{21}$$

where $o_k^{(l)}$ is the output of k -th neuron in (l) layer, $\varphi_k^{(l)}(\cdot)$ is the activation function, $w_{i,k}^{(l-1)}$ is the weight associated to the connection between i -th neuron from $(l-1)$ layer and k -th neuron from (l) layer, and $b_k^{(l)}$ is the bias parameter. We note that:

$$o_1^{(1)} \equiv x, \tag{22}$$

$$o_1^{(4)} \equiv H(x, \xi). \tag{23}$$

(iii) To reproduce the form of Eq. (11) we use the linear activation function for the neuron in the output layer:

$$\varphi_1^{(4)}(\cdot) = (\cdot), \tag{24}$$

where (\cdot) denotes the function argument, and we introduce the ξ -dependence via weights associated to the connections linking the last hidden layer with the output layer:

$$w_{i,1}^{(3)} = \xi^{2(i-1)}. \tag{25}$$

We also set the bias in the output neuron to:

$$b_1^{(4)} = 0. \tag{26}$$

$$\Phi_k^{(2)}(n) = \int_{-1}^1 dx x^n o_k^{(2)} = \int_{-1}^1 dx x^n \varphi_k^{(2)}(b_k^{(2)} + w_{1,k}^{(1)}x) \tag{30}$$

for the sigmoid function Eq. (27):

$$= (-1)^{n+1} \sum_{i=0}^n (w_{1,k}^{(1)})^{-i-1} \frac{\Gamma(1+n)}{\Gamma(1+n-i)} \left((-1)^{n+i} \text{Li}_{i+1} \left(-e^{b_k^{(2)} + w_{1,k}^{(1)}} \right) - \text{Li}_{i+1} \left(-e^{b_k^{(2)} - w_{1,k}^{(1)}} \right) \right), \tag{31}$$

or, for the ReLU function Eq. (28):

$$= \begin{cases} \frac{((-1)^n + 1)}{n+1} b_k^{(2)} - \frac{((-1)^n - 1)}{n+2} w_{1,k}^{(1)} & \text{for } \left| \frac{b_k^{(2)}}{w_{1,k}^{(1)}} \right| \geq 1 \text{ and } s_{b/w} \neq s_w, \\ 0 & \text{for } \left| \frac{b_k^{(2)}}{w_{1,k}^{(1)}} \right| \geq 1 \text{ and } s_{b/w} = s_w, \\ \frac{b_k^{(2)} w_{1,k}^{(1)} \left(-\frac{b_k^{(2)}}{w_{1,k}^{(1)}} \right)^{n+1} + s_w^{n+1} w_{1,k}^{(1)} \left(w_{1,k}^{(1)}(n+1) + s_w b_k^{(2)}(n+2) \right)}{(n+1)(n+2)w_{1,k}^{(1)}} & \text{otherwise,} \end{cases} \tag{32}$$

This choice means that $f_j(x)$ associated to ξ^j , see Eq. (11), is effectively described by an ANN with a single hidden layer corresponding to the first hidden layer of the full network, and the output layer corresponding to one neuron in the second hidden layer of the full network.

We are free to choose the activation function associated to the neurons in the hidden layers. For the purpose of further demonstration in the first hidden layer we use either the sigmoid function,

$$\varphi_k^{(2)}(\cdot) = \frac{1}{1 + \exp(-(\cdot))}, \tag{27}$$

or the rectifier function (ReLU),

$$\varphi_k^{(2)}(\cdot) = (\cdot) \Theta(\cdot), \tag{28}$$

where $\Theta(\cdot)$ is the Heaviside step function. In the second hidden layer the linear activation function will be used,

$$\varphi_k^{(3)}(\cdot) = (\cdot). \tag{29}$$

This choice allows us to analytically evaluate the Mellin coefficients using:

where:

$$s_{b/w} = \text{sgn} \left(\frac{b_k^{(2)}}{w_{1,k}^{(1)}} \right), \quad s_w = \text{sgn} \left(w_{1,k}^{(1)} \right), \tag{33}$$

independently of the choice of the activation function in the first hidden layer:

$$\begin{aligned} \Phi_k^{(3)}(n) &= \int_{-1}^1 dx x^n o_k^{(3)} \\ &= \int_{-1}^1 dx x^n \varphi_k^{(3)} \left(b_k^{(3)} + \sum_i w_{i,k}^{(2)} \varphi_i^{(2)} \left(w_{1,i}^{(1)} x + b_i^{(2)} \right) \right) \\ &= \text{mod}(n+1, 2) \frac{2}{n+1} b_k^{(3)} + \sum_i w_{i,k}^{(2)} \Phi_i^{(2)}(n), \end{aligned} \tag{34}$$

where $\text{Li}_i(x)$ is the polylogarithm and

$$\Phi_{1+j/2}^{(3)}(n) \equiv A_{n,j}. \tag{35}$$

These coefficients depend linearly on $b_k^{(3)}$ and $w_{i,k}^{(2)}$, so the latter can be found from a number of Mellin moments by solving the system of equations similar to that of Eq. (18). Here, $\mathbf{A}_j = (A_{0,j}, \dots, A_{i,j}, \dots, A_{N,j}, 0, 0)^T$ is the same as for the polynomial basis, while $\mathbf{w}_j =$

$$\left(b_{1+j/2}^{(3)}, w_{1,1+j/2}^{(2)}, \dots, w_{N+2,1+j/2}^{(2)} \right)^T \text{ and}$$

$$c_{n,i} = \begin{cases} 2 \bmod(n+1, 2)/(n+1) & n \leq N \text{ and } i = 0, \\ \Phi_i^{(2)}(n) & n \leq N \text{ and } i > 0, \\ 1 & n > N \text{ and } i = 0, \\ \varphi_i^{(2)} \left(b_i^{(2)} + (-1)^n w_{1,i}^{(1)} \right) & \text{otherwise.} \end{cases} \quad (36)$$

The size of the last hidden layer (third layer of the network: (3)) is $N/2 + 1$, while to ensure both the polynomiality and vanishing at $|x| = 1$ the size of the first hidden layer (second layer of the network: (2)) must be at least $N + 2$.

In our approach the coefficients of the first hidden layer, $w_{1,i}^{(1)}$ and $b_i^{(2)}$, are not constrained by the Mellin moments. Values of these coefficients can be taken random, or they can be fixed with further criteria, e.g. to obtain the best possible agreement between the input PDF and the resulting GPD in x -space. The sensitivity on the choice of $w_{1,i}^{(1)}$ and $b_i^{(2)}$ coefficients becomes small for large N .

3.4 Demonstration of models

We illustrate the modelling presented in this section with the help of the following pion GPD [59,60]:

$$H_\pi(x, \xi) = \Theta(x - |\xi|) \frac{30(1-x)^2(x^2 - \xi^2)}{(1 - \xi^2)^2} + \Theta(|\xi| - |x|) \frac{15(1-x)(\xi^2 - x^2)(x + 2x\xi + \xi^2)}{2\xi^3(1 + \xi)^2}. \quad (37)$$

We fix the truncation parameter N to either 4 or 8, and we evaluate the corresponding $A_{n,j}$ coefficients from $H_\pi(x, \xi)$. We use those coefficients to reconstruct $H_\pi(x, \xi)$ using either the polynomial or ANN basis (with either the sigmoid or ReLU activation functions). This way we check how well one may reconstruct the underlying GPD only knowing a number of its Mellin moments. The result is shown in Fig. 3 for $\xi = 0$ and $\xi = 0.5$.

In this Article we only show the most elementary use of the modelling based on Eq. (11). However several possible extensions or modifications are possible. In particular, one may extend the sum in Eq. (14) (for polynomial basis) or the number of neurons in the first hidden layer (for ANN basis) to introduce more free parameters that can be “spent” for a better description of the x -dependence and of higher Mellin coefficients for a given power of ξ . One can also try the modelling with an explicit PDF contribution:

$$H(x, \xi) = q(x) + \sum_{\substack{j=2 \\ \text{even}}}^N f_j(x)\xi^j, \quad (38)$$

which automatically fixes all $A_{n,0}$ coefficients. If needed, vanishing at $|x| = |\xi|$ can be enforced in this way:

$$H(x, \xi) = (x^2 - \xi^2) \sum_{\substack{j=0 \\ \text{even}}}^N f_j(x)\xi^j, \quad (39)$$

which typically causes an oscillatory behaviour along the x -axis. One may also introduce a separate D-term [44]:

$$H(x, \xi) = \text{sgn}(\xi)D(x/\xi) + \sum_{\substack{j=0 \\ \text{even}}}^N f_j(x)\xi^j, \quad (40)$$

here denoted by $D(x/\xi)$, which will contribute to $A_{n,n+1}$ coefficients. This separate contribution allows to reproduce GPDs, which are not smooth functions and usually exhibit a kink at $|x| = |\xi|$.

In our simple demonstration we do not check if the models fulfil the positivity constraint. With the pion model given by Eq. (37) used as a benchmark it is particularly difficult. From Eq. (3) we see that if the PDF vanishes for $x < 0$, the corresponding GPD must vanish as well for $x < -|\xi|$. We are not able to achieve this behaviour easily via our modelling of Mellin moments. However, we note that with a different benchmark model it should be possible to use a numerical enforcement of positivity. This method will be introduced in the next section of this Article.

4 Modelling in (β, α) -space

4.1 Principles of modelling

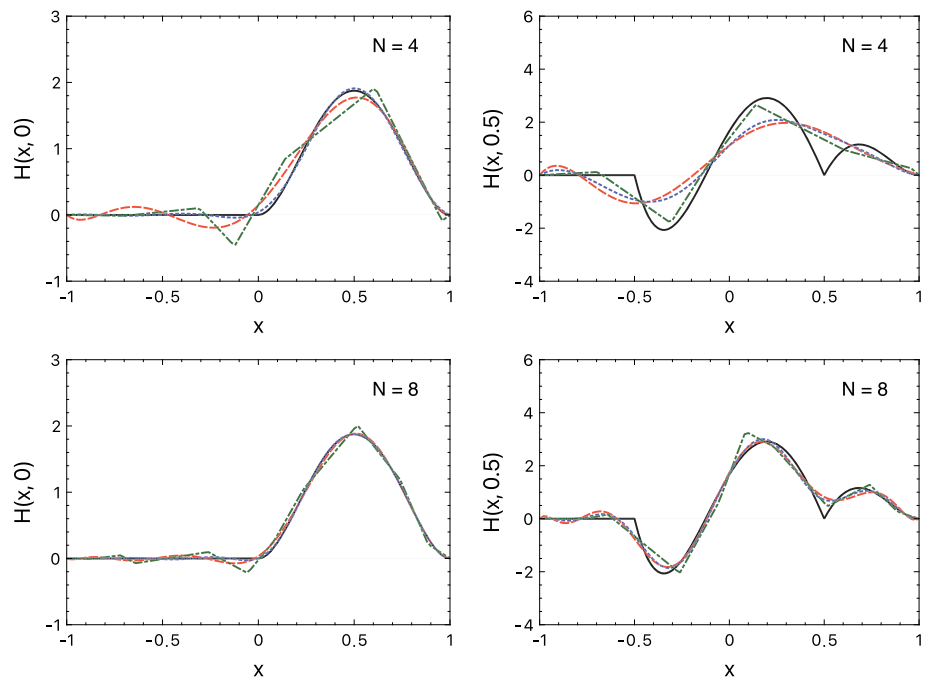
The advantage of modelling GPDs using double distributions is a natural provision of polynomiality by the Radon transform. In addition, a proper reduction to PDFs can be achieved easily, with a welcome possibility of keeping the singularity at $x = 0$ for only $\xi = 0$. The GPD support $x \in [-1, +1]$ is also naturally encoded in the double distribution support $|\alpha| + |\beta| \leq 1$.

We are modelling the GPD $H(x, \xi)$ with three contributions, each one being the subject to the Radon transform shown in Eq. (10):

$$(1 - x^2)F_C(\beta, \alpha) + (x^2 - \xi^2)F_S(\beta, \alpha) + \xi F_D(\beta, \alpha) \quad (41)$$

The first term, $(1 - x^2)F_C(\beta, \alpha)$ gives the “classical” contribution reproducing both the forward limit and the cross-over line $\xi = x$ [62]. This term is crucial for the comparison of GPD models with experimental data available today. The second term, $(x^2 - \xi^2)F_S(\beta, \alpha)$, gives a “shadow” contribution, which vanishes for both $\xi = 0$ and $\xi = x$ [24]. As we will see, its inclusion is important for a proper estimation of model uncertainties when GPDs are constrained by a sparse set of data. The third term, $\xi F_D(\beta, \alpha)$, only contributes to

Fig. 3 The pion model [59,60] (black solid line) given by Eq. (37) reconstructed from either five ($N = 4$, top row) or nine ($N = 8$, bottom row) of its first Mellin moments as a function of x for either $\xi = 0$ (left column) or $\xi = 0.5$ (right column). The reconstruction is done using a polynomial basis (red dashed line) or an ANN basis with either sigmoid activation function (blue dotted line) or ReLU activation function (green dash-dotted line)



the D-term, which is accessible in analyses of amplitudes for exclusive processes (see e.g. Ref. [37]). The inclusion of this term gives an extra flexibility to the model required to reproduce all x -moments of GPDs.

We note that for singlet GPDs, for which even Mellin moments vanish, the prefactors x^2 and ξ^2 in Eq. (41) do not break the polynomiality property. Indeed, for odd Mellin moments of order n , double distributions being the subject to the Radon transform shown in Eq. (10) give contributions up to the $A_{n,n-1}$ coefficient only – since the $A_{n,n}$ coefficient is zero for parity reasons. Therefore, adding either an x^2 or ξ^2 factor only brings contributions up to $A_{n,n+1}$, which is compatible with polynomiality.

4.2 $F_C(\beta, \alpha)$ contribution

We factorise $F_C(\beta, \alpha)$ into the forward limit, $f(\beta)$, which is well-known, at least for the GPD $H(x, \xi)$, and the profile function $h_C(\beta, \alpha)$:

$$F_C(\beta, \alpha) = f(\beta)h_C(\beta, \alpha) \frac{1}{1 - \beta^2}. \tag{42}$$

The prefactor $(1 - x^2)/(1 - \beta^2)$ arising from Eqs. (41) and (42) turns out to be convenient to preserve positivity with an ANN-based modelling of $h_C(\beta, \alpha)$. Since in this study we are only interested in the singlet contribution we fix:

$$f(\beta) = \text{sgn}(\beta)q(|\beta|). \tag{43}$$

The profile function must fulfil the following set of properties to ensure the correct behaviour of the GPD in (x, ξ) -space:

- (i) even parity w.r.t. the β variable,

$$h_C(\beta, \alpha) = h_C(-\beta, \alpha), \tag{44}$$

to keep the whole GPD an odd function of x .

- (ii) even parity w.r.t. the α variable,

$$h_C(\beta, \alpha) = h_C(\beta, -\alpha), \tag{45}$$

to keep the whole GPD an even function of ξ , and therefore to hold time reversal symmetry.

- (iii) the following normalisation:

$$\int_{-1+\beta}^{1-\beta} d\alpha h_C(\beta, \alpha) = 1, \tag{46}$$

to ensure the proper reduction to the PDF at $\xi = 0$.

- (iv) vanishing at the edge of the support region,

$$h_C(\beta, \alpha) = 0 \text{ for } |\beta| + |\alpha| = 1, \tag{47}$$

to avoid any singular behaviour, except at the $x = \xi = 0$ point, and to help enforcing the positivity property at $x \approx 1$.

We use the following model for $h_C(\beta, \alpha)$ fulfilling all the aforementioned requirements:

$$h_C(\beta, \alpha) = \frac{\text{ANN}_C(|\beta|, \alpha)}{\int_{-1+|\beta|}^{1-|\beta|} d\alpha \text{ANN}_C(|\beta|, \alpha)}. \tag{48}$$

Here, $\text{ANN}_C(|\beta|, \alpha)$ is the output of a single artificial neural network. We have chosen a feed forward artificial neural network with a single hidden layer, whose example is shown in Fig. 4. Our choice simplifies the evaluation of the integral in the denominator of Eq. (48). The signal is processed by the network in a way similar to the case described in Sect. 3.3. Keeping the nomenclature consistent throughout the text, two neurons in the input layer distribute the $|\beta|$ and α values:

$$o_1^{(1)} \equiv |\beta|, \tag{49}$$

$$o_2^{(1)} \equiv \alpha, \tag{50}$$

while the neuron in the output layer gives:

$$o_1^{(3)} \equiv \text{ANN}_C(|\beta|, \alpha). \tag{51}$$

A number of neurons in the hidden layer process information according to this equation:

$$o_k^{(2)} = \left[\varphi_k^{(2)} \left(b_k^{(2)} + w_{1,k}^{(1)}|\beta| + w_{2,k}^{(1)}\alpha / (1 - |\beta|) \right) - \varphi_k^{(2)} \left(b_k^{(2)} + w_{1,k}^{(1)}|\beta| + w_{2,k}^{(1)} \right) \right] + \left[w_{2,k}^{(1)} \rightarrow -w_{2,k}^{(1)} \right], \tag{52}$$

where the biases, $b_k^{(2)}$, and weights, $w_{1,k}^{(1)}$, $w_{2,k}^{(1)}$, are free parameters of the network, and where $\varphi_k^{(2)}(\cdot)$ is the activation function to be fixed by the network architect. For instance, it can be the sigmoid function from Eq. (27) or the ReLU function from Eq. (28). For the output layer we have:

$$o_1^{(3)} = \sum_k w_{k,1}^{(2)} o_k^{(2)}, \tag{53}$$

where we explicitly used the linear activation function. There is no bias parameter and the weights, $w_{k,1}^{(2)}$ are the other free parameters of the network. With the sigmoid function for $\varphi_k^{(2)}(\cdot)$, the normalisation factor is:

$$\int_{-1+|\beta|}^{1-|\beta|} d\alpha \text{ANN}_C(|\beta|, \alpha) = 2 \sum_k \frac{w_{k,1}^{(2)}}{w_{2,k}^{(1)}} (|\beta| - 1) \times \left[\log \left(\cosh \left(b_k^{(2)} - w_{2,k}^{(1)} + |\beta|w_{1,k}^{(1)} \right) \right) - \log \left(\cosh \left(b_k^{(2)} + w_{2,k}^{(1)} + |\beta|w_{1,k}^{(1)} \right) \right) + w_{2,k}^{(1)} \tanh \left(b_k^{(2)} - w_{2,k}^{(1)} + |\beta|w_{1,k}^{(1)} \right) + w_{2,k}^{(1)} \tanh \left(b_k^{(2)} + w_{2,k}^{(1)} + |\beta|w_{1,k}^{(1)} \right) \right] \tag{54}$$

4.3 $F_S(\beta, \alpha)$ contribution

The shadow contribution is modelled in a quite similar way to $F_C(\beta, \alpha)$:

$$F_S(\beta, \alpha) = f(\beta)h_S(\beta, \alpha). \tag{55}$$

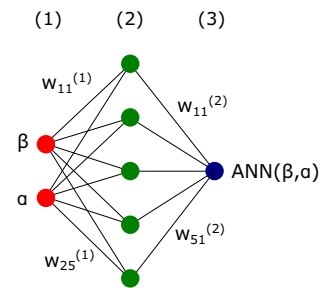


Fig. 4 Scheme of an exemplary artificial neural network used to represent the profile function in the double distribution model

Here, $f(\beta)$ has been already given in Eq. (43). The inclusion of this factor may come as a surprise, as the shadow GPD vanishes in the forward limit, but $f(\beta)$ helps here to fulfil the positivity constraint. Similarly to $h_C(\beta, \alpha)$, $h_S(\beta, \alpha)$ must be an even function with respect to both β and α variables, and we make it to vanish at $|\alpha| + |\beta| = 1$. Its normalisation is however different. We require

$$\int_{-1+|\beta|}^{1-|\beta|} d\alpha h_S(\beta, \alpha) = 0, \tag{56}$$

to ensure the vanishing at $\xi = 0$. We use the following model for $h_S(\beta, \alpha)$, which fulfils all the aforementioned requirements:

$$h_S(\beta, \alpha) / N_S = \frac{\text{ANN}_S(|\beta|, \alpha)}{\int_{-1+|\beta|}^{1-|\beta|} d\alpha \text{ANN}_S(|\beta|, \alpha)} - \frac{\text{ANN}_{S'}(|\beta|, \alpha)}{\int_{-1+|\beta|}^{1-|\beta|} d\alpha \text{ANN}_{S'}(|\beta|, \alpha)}. \tag{57}$$

Here, N_S is a scaling parameter, while $\text{ANN}_S(|\beta|, \alpha)$ and $\text{ANN}_{S'}(|\beta|, \alpha)$ are neural networks. As one may conclude, effectively, we are constructing the shadow contribution by subtracting two GPDs having the same forward limit. One of these GPDs, related to $\text{ANN}_S(|\beta|, \alpha)$, will be the subject of modelling. The other one, related to $\text{ANN}_{S'}(|\beta|, \alpha)$, provides an arbitrary reference point. For simplicity we take:

$$\text{ANN}_{S'}(|\beta|, \alpha) \equiv \text{ANN}_C(|\beta|, \alpha), \tag{58}$$

that is, one of two networks used in $F_S(\beta, \alpha)$ is the same as that used in $F_C(\beta, \alpha)$. In fits presented in Sect. 4.6 $\text{ANN}_S(|\beta|, \alpha)$ will have the same architecture as $\text{ANN}_{S'}(|\beta|, \alpha)$, but different free parameters (weights and biases).

4.4 $F_D(\beta, \alpha)$ contribution

This contribution gives the D-term:

$$F_D(\beta, \alpha) = \delta(\beta)D(\alpha), \tag{59}$$

where:

$$D(\alpha) = (1 - \alpha^2) \sum_{\substack{i=1 \\ \text{odd}}}^N d_i C_i^{3/2}(\alpha), \quad (60)$$

d_i are the coefficients of the expansion on the family of Gegenbauer polynomials and N is a truncation parameter. This definition in (x, ξ) -space explicitly gives:

$$D(z) = \Theta(1 - |z|)(1 - z^2) \sum_{\substack{i=1 \\ \text{odd}}}^N d_i C_i^{3/2}(z), \quad (61)$$

where $z = x/\xi$.

4.5 Remarks

In Eq. (52) the variable α is scaled by $1 - |\beta|$, such that $\alpha/(1 - |\beta|)$ spans over the range of $(-1, 1)$. Such a standardisation of variables is typical for neural networks, as it leads to a faster convergence and it allows to describe equally well the dependencies on all input variables.

The factorisation expressed by Eq. (42) is arbitrary. In principle, the neural network could absorb some, if not all, information about PDFs. For instance, we could express a double distribution F in the following way:

$$F(\beta, \alpha) = \text{sgn}(\beta) |\beta|^{-\delta} (1 - |\beta|)^\gamma \text{ANN}(\beta, \alpha), \quad (62)$$

where δ and γ are powers driving the behaviour of PDF at $x \rightarrow 0$ and $x \rightarrow 1$, respectively, and where the normalisation factor, as used in the denominator of Eq. (48), is not needed anymore. This kind of factorisation may be useful when (semi-)inclusive and exclusive data will be simultaneously used to constrain GPDs.

4.6 Fits to pseudo-data

We now discuss technicalities of constraining GPD models defined in (β, α) -space from data. We first elaborate on aspects of our numerical analysis like minimisation, estimation of uncertainties, treatment of outliers, regularisation and positivity enforcement. Then, we present three examples demonstrating possible applications of our approach. We particularly emphasise the estimation of uncertainties in unconstrained kinematic domains. As in the rest of this Article, we will only focus on the GPD $H(x, \xi)$, neglecting the t and μ_F^2 -dependencies. We note that the actual extraction of GPDs from experimental data will require taking into account both the t -dependence and the evolution equations, but also other types of GPDs. This is beyond the scope of this Article, where we only pave the way towards such analysis. The complex link between the GPD $H(x, \xi)$ and experimental data is also the reason why we do not emulate experimental uncertainties. With no such uncertainties taken into account our

demonstration may be seen as a level-0 closure test known from the PDF methodology [39, 63, 64].

In our numerical analysis the minimisation, i.e. the procedure of constraining free parameters, is done with the genetic algorithm [65]. This algorithm is iterative. In a single iteration many sets of free parameters, referred to in the literature as “candidates”, are simultaneously checked against a “fitness function”. After evaluation, the best candidates, i.e. those characterised by the best values of the fitness function, are “crossed over” with the hope of obtaining even better candidates to be used in the next iteration. The cross-over is followed by a “mutation”, where a number of free parameters is randomly changed, allowing for a significant reduction of the risk of converging to a local minimum. We note that, since in a given iteration of the genetic algorithm the fitness function is simultaneously evaluated for all candidates, multithreading computing can be employed to improve the performance of the minimisation.

The employed fitness function is the root mean squared relative error [66]:

$$\text{RMSRE} = \sqrt{\frac{1}{N_{\text{pts}}} \sum_{i=1}^{N_{\text{pts}}} \left(\frac{H_0(x_i, \xi_i) - H(x_i, \xi_i)}{H_0(x_i, \xi_i)} \right)^2}. \quad (63)$$

Here, the sum runs over N_{pts} points probing (x, ξ) phase-space, while $H_0(x_i, \xi_i)$ and $H(x_i, \xi_i)$ are GPDs given by the reference (here GK) and ANN models, respectively. The RMSRE allows to avoid significant differences of contributions to the fitness function coming from various domains of (x, ξ) -space.

The pseudo-data used in this analysis do not have uncertainties. However, we are still interested to know what are the uncertainties of models in domains unconstrained by data, i.e. what is the effect of the sparsity of data, in particular when only $x = \xi$ data are used in the fit and the neural network was designed to explore the whole $x \neq \xi$ domain. To estimate this kind of uncertainties we repeat the minimization multiple times, each time starting the genetic algorithms with a random set of initial parameters. We refer to the outcome of such a single minimization as “a replica”. A replica is a possible realisation of the GPD model. All replicas reproduce the data used in the minimization, but because of the flexibility of ANNs their behaviour in unconstrained kinematic domains can be considered random. We note that this randomness can be unintentionally suppressed if one uses too small networks or restricts too much the values of weights and biases.

We use the spread of 100 replicas to estimate the model uncertainty in a given kinematic point. Because of the complexity of our fits and unavoidable problems in the minimisation, like falling into local minima, it sometimes happens that a single replica gives values that are very different compared

to those given by other replicas. This problem is typical in phenomenological studies (see e.g. Ref. [38]), and requires finding and removing the problematic replicas, or suppressing them in the estimation of uncertainties. In this analysis we employ the following algorithm: for a given population of replicas (i) we randomly choose 1000 points of (x, ξ) , (ii) for a given point we evaluate the GPD values that the replicas give in that point, and then we check which of those values are outliers by applying the 3σ -rule [67], (iii) the replicas giving values identified as outliers in more than 10% points are entirely removed from the estimation of uncertainties.

As in other analyses using effectively nonparametric models, a regularisation must be used to avoid biased results caused by over-fitting. Without a regularisation, an ANN tends to describe the fitted data extremely precisely, resulting in a minimal variance evaluated for these data. It does not mean though that the ANN will describe equally well any other data, i.e. that it will have a predictive power. In general, a bias may appear, because too much attention was paid to describe the fitted data, and the ANN does not describe well the general trends. Many types of regularisation methods exist, and the selection of a given method typically depends on the problem being under consideration. In this analysis we use the so-called dropout method [68]. In this method in a given iteration of the minimisation algorithm (here: the genetic algorithm) a predefined fraction of neurons (here: 10%) is randomly dropped, i.e. some neurons become inactive, they do not process the signal. The output of other neurons is correspondingly scaled to compensate for the loss. Effectively, in each iteration a different subset of neurons in the hidden layer is used, which avoids focusing on details only characterising the fitted data. The fraction of dropped neurons we use has been validated against independent samples with respect to those used in the fits presented in the following of this section.

The genetic algorithm allows for a straightforward implementation of penalisation methods to avoid unwanted results, in particular those not fulfilling binary-like conditions, like inequalities. We use this feature to enforce the positivity condition. To achieve this, for each set of free parameters (i.e. for each candidate) we check by how much we can scale $F_S(\beta, \alpha)$ by changing the N_S parameter to saturate the condition given by Eq. (3). We check this in 10,000 points covering the $(x, \xi < x)$ domain. If we are not able to scale $F_S(\beta, \alpha)$ to respect the constraint in all of those points, the candidate is discarded, i.e. it is dropped from the population of candidates. If we can, we scale $F_S(\beta, \alpha)$ accordingly, i.e. we take the highest allowed value of $|N_S|$. This simple method of enforcing the positivity constraints requires a substantial computing power, but gives satisfactory results. We conclude that up to the numerical noise the constraint is fulfilled, which we demonstrate with Fig. 10 showing the x -dependencies of replicas with respect to the positivity bound for few val-

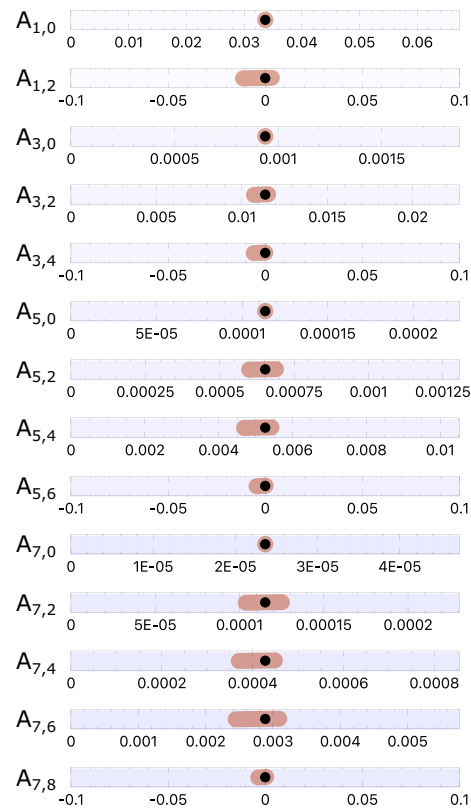


Fig. 5 Result of constraining the ANN model using 400 points evaluated with the GK model [25–27] for sea quarks: $x \neq \xi$ case, positivity not enforced, all three contributions ($F_C(\beta, \alpha)$, $F_S(\beta, \alpha)$, $F_D(\beta, \alpha)$) used in the presentation of results. The plot shows the first Mellin coefficient for odd moments (even moments strictly vanish) evaluated for the GK (black dots) and ANN (solid bands) models

ues of ξ . We note that after the minimisation, $F_C(\beta, \alpha)$ and $F_S(\beta, \alpha)$ may not fulfill alone the positivity constraint given by the chosen PDF, but they do in sum. It indicates that the inclusion of $F_S(\beta, \alpha)$ gives an extra freedom to achieve the positivity.

We now show and discuss results of three extractions of GPDs with pseudo-data generated with the GK model [25–27]. In all three cases $\text{ANN}_C(|\beta|, \alpha)$ used in both $F_C(\beta, \alpha)$ and $F_S(\beta, \alpha)$, and $\text{ANN}_S(|\beta|, \alpha)$ used in $F_S(\beta, \alpha)$ consist of 5 neurons in the hidden layer, each, and we use the sigmoid activation function. The $F_D(\beta, \alpha)$ contribution consists of 5 elements, i.e. the sum in Eq. (60) runs up to $i = 9$.

In the first test-case we train our model on $N_{\text{pts}} = 400$ points uniformly covering the domain of $-4 < \log_{10}(x) < \log_{10}(0.95)$, $-4 < \log_{10}(\xi) < \log_{10}(0.95)$, $t = 0$ and $Q^2 = 4 \text{ GeV}^2$. That is, in this scenario, $x \neq \xi$ data are used to constrain our ANN model. The purpose of this test is to check if our approach can be used to reproduce GPD models such as GK. All three contributions from Eq. (41) are used in the fit, even if we *a priori* know that GK does not include any D-term. It is because $F_C(\beta, \alpha)$ and $F_S(\beta, \alpha)$ may

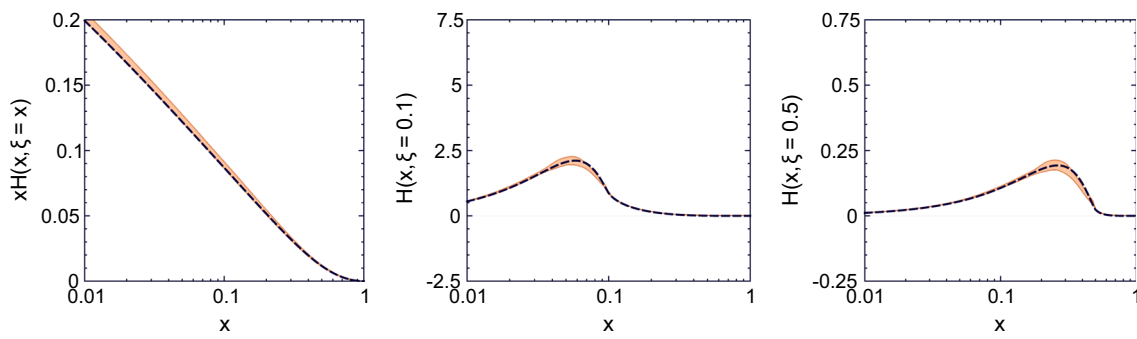


Fig. 6 Result of constraining the ANN model using 400 points evaluated with the GK model [25–27] for sea quarks: $x \neq \xi$ case, positivity not enforced, all three contributions ($F_C(\beta, \alpha)$, $F_S(\beta, \alpha)$, $F_D(\beta, \alpha)$) shown. See the text for more details. The comparison is for (left) $\xi = x$, (center) $\xi = 0.1$ and (right) $\xi = 0.5$. The dashed lines denote the GK model, while the bands represent the result of the fit in the form of a 68% confidence level

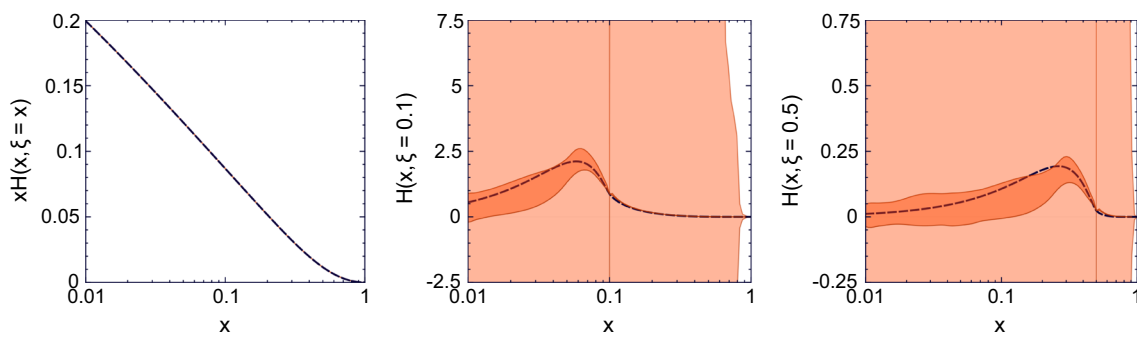


Fig. 7 Result of constraining the ANN model using 200 points evaluated with GK model [25–27] for sea quarks: $x = \xi$ case, positivity not enforced, $F_D(\beta, \alpha)$ not shown. See the text for more details. The comparison is for (left) $x = \xi$, (center) $\xi = 0.1$ and (right) $\xi = 0.5$. The dashed lines denote the GK model, while the bands represent the result of the fit in the form of a 68% confidence level. The inner bands show $F_C(\beta, \alpha)$ contribution, alone, while the outer bands are for $F_C(\beta, \alpha) + F_S(\beta, \alpha)$

contribute to the D-term due to the $(1 - x^2)$ and $(x^2 - \xi^2)$ factors, respectively. With $F_D(\beta, \alpha)$ included in the fit we are able to compensate for those contributions. We do not enforce the positivity constraint, as the GK model violates the inequality shown in Eq. (3). Enforcing this positivity inequality would lead to a discrepancy between our model and GK. We note that in average the fit of a single replica ends with a $\text{RMSRE} \approx 0.016$ and takes ≈ 1.9 h on 40 computing cores. Such a small RMSRE indicates that our model can firmly reproduce the GK model in the probed domain of (x, ξ) -space. A higher number of neurons in the hidden layers would lead to an even smaller RMSRE. The graphical representation of the results is shown in Fig. 5 for Mellin moments and in Fig. 6 for the x -dependence. As highlighted by the low average value of the RMSRE, the agreement between the fitted ANN model and GK is very good, hence we conclude to a successful outcome of the whole test.

In the second test-case only $x = \xi$ data are used to constrain our GPD model. These are $N_{\text{pts}} = 200$ points uniformly covering the domain of $-4 < \log_{10}(x = \xi) <$

$\log_{10}(0.95)$, $t = 0$ and $Q^2 = 4 \text{ GeV}^2$. The purpose of this test is to demonstrate how our approach can be used to reconstruct GPDs from amplitudes for processes like DVCS and TCS, when described at LO. The positivity inequality is not enforced. The result is shown in Fig. 7. Since we are not interested in the D-term (it is not constrained by $x = \xi$ data, but can be accessed elsewhere, namely in dispersive analyses) the contribution coming from $F_D(\beta, \alpha)$ is removed from the presentation of results. The contribution to the D-term generated by $F_C(\beta, \alpha)$ and $F_S(\beta, \alpha)$ is small with respect to the uncertainties and will be discussed in more detail in the next paragraph. The fit of a single replica ends in this test-case in average with a $\text{RMSRE} \approx 0.0019$ and takes ≈ 0.5 h on 40 computing cores. In Fig. 7 one may observe exploding uncertainties, except for the $\xi = x$ line. This behaviour is expected. From the same figure one may judge on the importance of including the shadow contribution, $F_S(\beta, \alpha)$. We see that $F_C(\beta, \alpha)$ is over-constrained by the necessity of reproducing the GPD at both $\xi = 0$ and $\xi = x$ lines. Since $h_C(\beta, \alpha)$ is normalized, see Eqs. (46) and (48), $\text{ANN}_C(|\beta|, \alpha)$ associ-

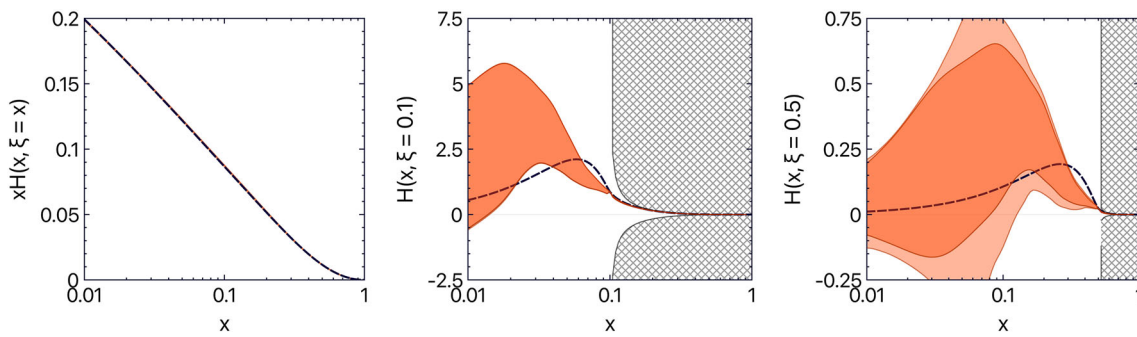


Fig. 8 Result of constraining the ANN model using 200 points evaluated with the GK model [25–27] for sea quarks: $x = \xi$ case, positivity enforced, $F_D(\beta, \alpha)$ not shown. See the text for more details. The comparison is for (left) $x = \xi$, (center) $\xi = 0.1$ and (right) $\xi = 0.5$. The dashed lines denote the GK model, while the bands represent

the result of the fit in the form of a 68% confidence level. The inner bands show $F_C(\beta, \alpha)$ contribution, alone, while the outer bands are for $F_C(\beta, \alpha) + F_S(\beta, \alpha)$. The regions excluded by the positivity constraint are denoted by the hatched bands

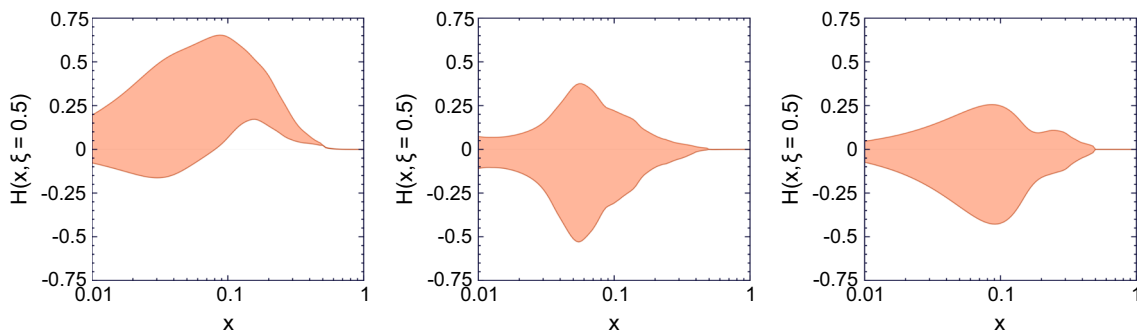


Fig. 9 Result of constraining the ANN model using 200 points evaluated with the GK model [25–27] for sea quarks: $x = \xi$ case, positivity enforced. Contributions coming from (left) $F_C(\beta, \alpha)$ and (center)

$F_S(\beta, \alpha)$ are shown. (right) Contribution to D-term induced by both $F_C(\beta, \alpha)$ and $F_S(\beta, \alpha)$ is shown. All plots are for $\xi = 0.5$

ated to this term lacks the flexibility allowing for a significant contribution to the uncertainties in kinematic domains that are unconstrained by data.

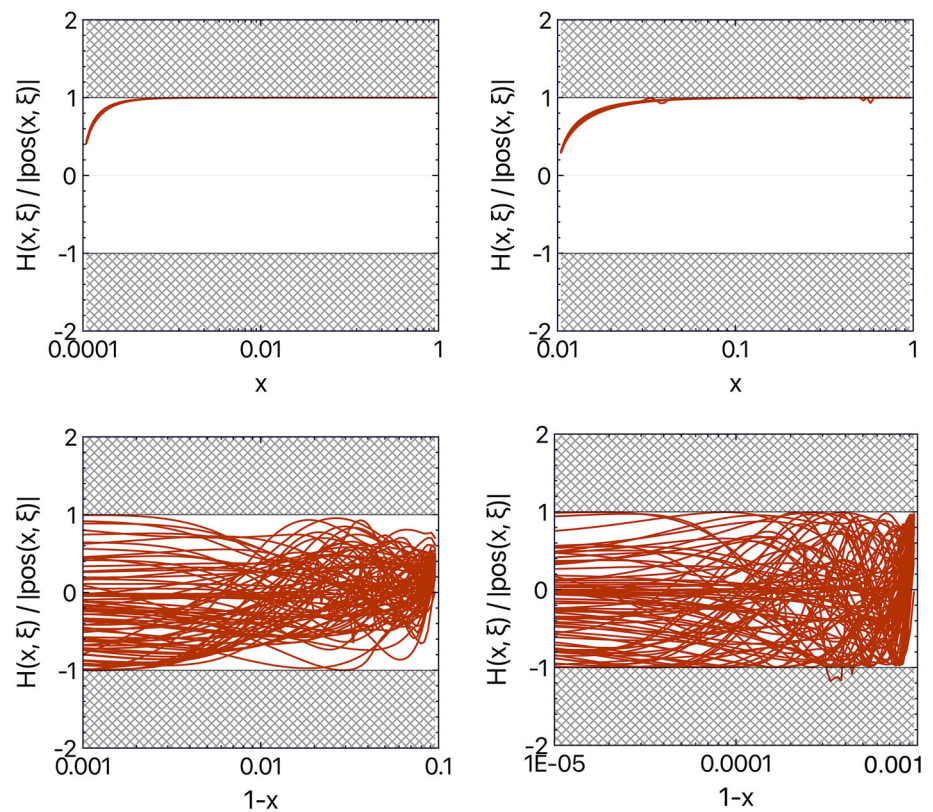
The conditions used in the third test-case are the same as in the second one, except now the positivity inequality is enforced to show its impact on the reduction of uncertainties. The algorithm for the selection of replicas fulfilling the positivity inequality introduces a substantial computational burden. The fit of a single replica in average ends with a $RMSRE \approx 0.0075$ and takes ≈ 26 h on 40 computing cores. One should be able to optimise the algorithm to make this time shorter. The result is shown in Fig. 8. We observe a significant reduction of uncertainties due to positivity. The contributions coming from the $F_C(\beta, \alpha)$ and $F_S(\beta, \alpha)$ terms, together with the D-term induced by those two terms, is shown in Fig. 9. Both contributions are substantial.

5 Summary

The lack of GPD flexible parameterisations fulfilling all required theoretical constraints has been an obstacle slowing down the completion of global fits to experimental data comparable to those currently achieved in the PDF community. Often GPD models are either too rigid to accommodate the measurements, or insufficiently constrained from the theoretical point of view. Since the different theoretical requirements are hardly met in common analytic expressions, neural networks are appealing tools to obtain flexible, yet complex, parameterisations. To the best of our knowledge, our present study is the first providing concrete elements to actually build GPD models based on neural network techniques.

In this Article we discuss different strategies to model GPDs in (x, ξ) and (β, α) spaces. In all cases GPD Mellin moments and polynomiality play a central role. We show in particular that a neural network description of double distributions is an efficient way to implement polynomiality,

Fig. 10 Result of constraining the ANN model using 200 points evaluated with the GK model [25–27] for sea quarks: $x = \xi$ case, positivity enforced. Distributions of GPDs normalised by the positivity bound are shown as a function of either x or $(1-x)$. Solid red curves are for replicas used in this analysis. The regions excluded by the positivity constraint are denoted by the hatched bands. The plots are for (top left) $\xi = 0.0001$, (top right) $\xi = 0.01$, (bottom left) $\xi = 0.9$ and (bottom right) $\xi = 0.999$



positivity, discrete symmetries, as well as the reduction to any freely chosen PDF in the forward limit.

An increased control of systematic effects is highly desirable in front of the precision era of GPD physics and extractions. To this aim we also address questions of a practical nature when dealing with many parameters: existence of local minima in optimisation routines, presence of outliers in statistical data analysis, etc.

In view of future GPD fits to experimental DVCS data, we pay a special attention to singlet GPDs, and to contributions with both vanishing forward limit and vanishing DVCS amplitude at leading order in perturbation theory. Our parameterisations may thus be used as tools to study the model dependency affecting extractions of GPDs and derived quantities, like GPD Mellin moments and the total angular momentum of hadrons. Since these parameterisations fulfil theory driven constraints, they may be conveniently used in current and future analyses of GPDs, or in connection e.g. to lattice QCD, either in x -space or through Mellin moments.

Acknowledgements The authors thank Valerio Bertone, Feliciano Carlos De Soto Borrero, Jose Manuel Morgado Chávez, Markus Diehl, Cedric Mezrag, Pepe Rodriguez-Quintero, Jorge Segovia and Jakub Wagner for valuable discussions. H.M. acknowledges the support by the Polish National Science Centre with the grant no. 2017/26/M/ST2/01074. P.S. is supported by the Polish National Science Centre with the grant no. 2019/35/D/ST2/00272. The computing resources of Świerk Computing Centre, Poland are greatly acknowledged. This project was supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No

824093. This work is supported in part in the framework of the GLUODYNAMICS project funded by the “P2IO LabEx (ANR-10-LABX-0038)” in the framework “Investissements d’Avenir” (ANR-11-IDEX-0003-01) managed by the Agence Nationale de la Recherche (ANR), France.

Data Availability Statement This manuscript has no associated data or the data will not be deposited. [Authors’ comment: All final data obtained in this study are contained in this published article.]

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.
Funded by SCOAP³.

References

1. D. Müller, D. Robaschik, B. Geyer, F.M. Dittes, J. Hořejši, *Fortsch. Phys.* **42**, 101 (1994). <https://doi.org/10.1002/prop.2190420202>
2. X.D. Ji, *Phys. Rev. Lett.* **78**, 610 (1997). <https://doi.org/10.1103/PhysRevLett.78.610>
3. X.D. Ji, *Phys. Rev. D* **55**, 7114 (1997). <https://doi.org/10.1103/PhysRevD.55.7114>

4. A.V. Radyushkin, Phys. Lett. B **385**, 333 (1996). [https://doi.org/10.1016/0370-2693\(96\)00844-1](https://doi.org/10.1016/0370-2693(96)00844-1)
5. A.V. Radyushkin, Phys. Rev. D **56**, 5524 (1997). <https://doi.org/10.1103/PhysRevD.56.5524>
6. M. Burkardt, Phys. Rev. D **62**, 119903 (2000). <https://doi.org/10.1103/PhysRevD.62.071503>. [Erratum: Phys. Rev. D 66, 119903 (2002)]
7. M. Burkardt, Int. J. Mod. Phys. A **18**, 173 (2003). <https://doi.org/10.1142/S0217751X03012370>
8. M. Burkardt, Phys. Lett. B **595**, 245 (2004). <https://doi.org/10.1016/j.physletb.2004.05.070>
9. X.D. Ji, W. Melnitchouk, X. Song, Phys. Rev. D **56**, 5511 (1997). <https://doi.org/10.1103/PhysRevD.56.5511>
10. M.V. Polyakov, A.G. Shuvaev, On “dual” parametrizations of generalized parton distributions (2002). [arXiv:hep-ph/0207153](https://arxiv.org/abs/hep-ph/0207153)
11. M.V. Polyakov, Phys. Lett. B **555**, 57 (2003). [https://doi.org/10.1016/S0370-2693\(03\)00036-4](https://doi.org/10.1016/S0370-2693(03)00036-4)
12. M.V. Polyakov, P. Schweitzer, Int. J. Mod. Phys. A **33**(26), 1830025 (2018). <https://doi.org/10.1142/S0217751X18300259>
13. C. Lorcé, H. Moutarde, A.P. Trawiński, Eur. Phys. J. C **79**(1), 89 (2019). <https://doi.org/10.1140/epjc/s10052-019-6572-3>
14. E.R. Berger, M. Diehl, B. Pire, Eur. Phys. J. C **23**, 675 (2002). <https://doi.org/10.1007/s100520200917>
15. J.C. Collins, L. Frankfurt, M. Strikman, Phys. Rev. D **56**, 2982 (1997). <https://doi.org/10.1103/PhysRevD.56.2982>
16. A. Accardi et al., Eur. Phys. J. A **52**(9), 268 (2016). <https://doi.org/10.1140/epja/i2016-16268-9>
17. R. Abdul Khalek, et al., Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report (2021). [arXiv:2103.05419](https://arxiv.org/abs/2103.05419) [physics.ins-det]
18. D.P. Anderle et al., Front. Phys. (Beijing) **16**(6), 64701 (2021). <https://doi.org/10.1007/s11467-021-1062-0>
19. J.L. Abelleira Fernandez et al., J. Phys. G **39**, 075001 (2012). <https://doi.org/10.1088/0954-3899/39/7/075001>
20. K. Kumericki, S. Liuti, H. Moutarde, Eur. Phys. J. A **52**(6), 157 (2016). <https://doi.org/10.1140/epja/i2016-16157-3>
21. N. d’Hose, S. Niccolai, A. Rostomyan, Eur. Phys. J. A **52**(6), 151 (2016). <https://doi.org/10.1140/epja/i2016-16151-9>
22. L. Favart, M. Guidal, T. Horn, P. Kroll, Eur. Phys. J. A **52**(6), 158 (2016). <https://doi.org/10.1140/epja/i2016-16158-2>
23. P. Chatagnon et al., Phys. Rev. Lett. **127**(26), 262501 (2021). <https://doi.org/10.1103/PhysRevLett.127.262501>
24. V. Bertone, H. Dutrieux, C. Mezrag, H. Moutarde, P. Sznajder, Phys. Rev. D **103**(11), 114019 (2021). <https://doi.org/10.1103/PhysRevD.103.114019>
25. S.V. Goloskokov, P. Kroll, Eur. Phys. J. C **42**, 281 (2005). <https://doi.org/10.1140/epjc/s2005-02298-5>
26. S.V. Goloskokov, P. Kroll, Eur. Phys. J. C **53**, 367 (2008). <https://doi.org/10.1140/epjc/s10052-007-0466-5>
27. S.V. Goloskokov, P. Kroll, Eur. Phys. J. C **65**, 137 (2010). <https://doi.org/10.1140/epjc/s10052-009-1178-9>
28. M. Vanderhaeghen, P.A.M. Guichon, M. Guidal, Phys. Rev. Lett. **80**, 5064 (1998). <https://doi.org/10.1103/PhysRevLett.80.5064>
29. M. Vanderhaeghen, P.A.M. Guichon, M. Guidal, Phys. Rev. D **60**, 094017 (1999). <https://doi.org/10.1103/PhysRevD.60.094017>
30. K. Goeke, M.V. Polyakov, M. Vanderhaeghen, Prog. Part. Nucl. Phys. **47**, 401 (2001). [https://doi.org/10.1016/S0146-6410\(01\)00158-2](https://doi.org/10.1016/S0146-6410(01)00158-2)
31. M. Guidal, M.V. Polyakov, A.V. Radyushkin, M. Vanderhaeghen, Phys. Rev. D **72**, 054013 (2005). <https://doi.org/10.1103/PhysRevD.72.054013>
32. V. Bertone and collaborators, in preparation (2021)
33. M. Constantinou, Eur. Phys. J. A **57**(2), 77 (2021). <https://doi.org/10.1140/epja/s10050-021-00353-7>
34. B. Berthou et al., Eur. Phys. J. C **78**(6), 478 (2018). <https://doi.org/10.1140/epjc/s10052-018-5948-0>
35. K. Kumericki, Status of the GeParD code (2021). Joint GDR-QCD/QCD@short distances and STRONG2020/PARTONS/FTE@LHC/NLOAccess workshop
36. K. Kumericki, Nature **570**(7759), E1 (2019). <https://doi.org/10.1038/s41586-019-1211-6>
37. H. Dutrieux, C. Lorcé, H. Moutarde, P. Sznajder, A. Trawiński, J. Wagner, Eur. Phys. J. C **81**(4), 300 (2021). <https://doi.org/10.1140/epjc/s10052-021-09069-w>
38. R.D. Ball et al., Nucl. Phys. B **867**, 244 (2013). <https://doi.org/10.1016/j.nuclphysb.2012.10.003>
39. R.D. Ball et al., JHEP **04**, 040 (2015). [https://doi.org/10.1007/JHEP04\(2015\)040](https://doi.org/10.1007/JHEP04(2015)040)
40. E.R. Nocera, R.D. Ball, S. Forte, G. Ridolfi, J. Rojo, Nucl. Phys. B **887**, 276 (2014). <https://doi.org/10.1016/j.nuclphysb.2014.08.008>
41. L. Del Debbio, T. Giani, J. Karpie, K. Orginos, A. Radyushkin, S. Zafeiropoulos, JHEP **02**, 138 (2021). [https://doi.org/10.1007/JHEP02\(2021\)138](https://doi.org/10.1007/JHEP02(2021)138)
42. X.D. Ji, J. Phys. G **24**, 1181 (1998). <https://doi.org/10.1088/0954-3899/24/7/002>
43. A.V. Radyushkin, Phys. Lett. B **449**, 81 (1999). [https://doi.org/10.1016/S0370-2693\(98\)01584-6](https://doi.org/10.1016/S0370-2693(98)01584-6)
44. M.V. Polyakov, C. Weiss, Phys. Rev. D **60**, 114017 (1999). <https://doi.org/10.1103/PhysRevD.60.114017>
45. A.V. Radyushkin, Phys. Rev. D **59**, 014030 (1999). <https://doi.org/10.1103/PhysRevD.59.014030>
46. B. Pire, J. Soffer, O. Teryaev, Eur. Phys. J. C **8**, 103 (1999). <https://doi.org/10.1007/s100529901063>
47. M. Diehl, T. Feldmann, R. Jakob, P. Kroll, Nucl. Phys. B **596**, 33 (2001). [https://doi.org/10.1016/S0550-3213\(00\)00684-2](https://doi.org/10.1016/S0550-3213(00)00684-2). [Erratum: Nucl. Phys. B 605, 647-647 (2001)]
48. P.V. Pobylitsa, Phys. Rev. D **65**, 077504 (2002). <https://doi.org/10.1103/PhysRevD.65.077504>
49. P.V. Pobylitsa, Phys. Rev. D **65**, 114015 (2002). <https://doi.org/10.1103/PhysRevD.65.114015>
50. P.V. Pobylitsa, Phys. Rev. D **66**, 094002 (2002). <https://doi.org/10.1103/PhysRevD.66.094002>
51. P.V. Pobylitsa, Phys. Rev. D **67**, 034009 (2003). <https://doi.org/10.1103/PhysRevD.67.034009>
52. P.V. Pobylitsa, Phys. Rev. D **67**, 094012 (2003). <https://doi.org/10.1103/PhysRevD.67.094012>
53. P.V. Pobylitsa, Phys. Rev. D **70**, 034004 (2004). <https://doi.org/10.1103/PhysRevD.70.034004>
54. M. Diehl, Phys. Rept. **388**, 41 (2003). <https://doi.org/10.1016/j.physrep.2003.08.002>
55. K. Kumericki, D. Mueller, Nucl. Phys. B **841**, 1 (2010). <https://doi.org/10.1016/j.nuclphysb.2010.07.015>
56. D. Müller, M.V. Polyakov, K.M. Semenov-Tian-Shansky, JHEP **03**, 052 (2015). [https://doi.org/10.1007/JHEP03\(2015\)052](https://doi.org/10.1007/JHEP03(2015)052)
57. S.V. Goloskokov, P. Kroll, Eur. Phys. J. C **50**, 829 (2007). <https://doi.org/10.1140/epjc/s10052-007-0228-4>
58. N. Chouika, C. Mezrag, H. Moutarde, J. Rodríguez-Quintero, Eur. Phys. J. C **77**(12), 906 (2017). <https://doi.org/10.1140/epjc/s10052-017-5465-6>
59. N. Chouika, C. Mezrag, H. Moutarde, J. Rodríguez-Quintero, Phys. Lett. B **780**, 287 (2018). <https://doi.org/10.1016/j.physletb.2018.02.070>
60. J.M.M. Chavez, V. Bertone, F.D.S. Borrero, M. Defurne, C. Mezrag, H. Moutarde, J. Rodríguez-Quintero, J. Segovia, Pion GPDs: A path toward phenomenology (2021). [arXiv:2110.06052](https://arxiv.org/abs/2110.06052) [hep-ph]
61. G. Cybenko, Mathematics of Control. Signals and Systems **2**(4), 303 (1989). <https://doi.org/10.1007/BF02551274>
62. I.V. Musatov, A.V. Radyushkin, Phys. Rev. D **61**, 074027 (2000). <https://doi.org/10.1103/PhysRevD.61.074027>
63. G. Watt, R.S. Thorne, JHEP **08**, 052 (2012). [https://doi.org/10.1007/JHEP08\(2012\)052](https://doi.org/10.1007/JHEP08(2012)052)

64. L. Del Debbio, T. Giani, M. Wilson, (2021). [arXiv:2111.05787](https://arxiv.org/abs/2111.05787) [hep-ph]
65. M. Mitchell, *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, MA, USA, 1998)
66. M. Göçken, M. Özçalıcı, A. Boru, A.T. Dosdoğru, *Expert Syst. Appl.* **44**, 320 (2016). <https://doi.org/10.1016/j.eswa.2015.09.029>
67. I.F. Ilyas, X. Chu, *Data Cleaning* (ACM, 2019). <https://doi.org/10.1145/3310205>
68. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *J. Mach. Learn. Res.* **15**(56), 1929 (2014)