Nucleon axial form factor from a Bayesian neural-network analysis of neutrino-scattering data

Luis Alvarez-Ruso,[1] Krzysztof M. Graczyk,[2] and Eduardo Saul-Sala[1]

[1]*Departamento de Física Teórica and Instituto de Física Corpuscular (IFIC), Centro Mixto UVEG-CSIC, Valencia, Spain*
[2]*Institute of Theoretical Physics, University of Wrocław, plac M. Borna 9, 50-204, Wrocław, Poland*

The Bayesian approach for feedforward neural networks has been applied to the extraction of the nucleon axial form factor from the neutrino-deuteron-scattering data measured by the Argonne National Laboratory bubble-chamber experiment. This framework allows to perform a model-independent determination of the axial form factor from data. When the low $0.05 < Q^2 < 0.10$-GeV^2 data are included in the analysis, the resulting axial radius disagrees with available determinations. Furthermore, a large sensitivity to the corrections from the deuteron structure is obtained. In turn, when the low-Q^2 region is not taken into account with or without deuteron corrections, no significant deviations from previous determinations have been observed. A more accurate determination of the nucleon axial form factor requires new precise measurements of neutrino-induced quasielastic scattering on hydrogen and deuterium.

DOI: [10.1103/PhysRevC.99.025204](https://doi.org/10.1103/PhysRevC.99.025204)

I. INTRODUCTION

A good understanding of neutrino interactions with matter are crucial to achieve the precision goals of oscillation experiments that aim at a precise determination of neutrino properties [1]. In current (T2K, NOvA) and future (DUNE, HyperK) oscillation experiments with few-GeV neutrinos, a realistic modeling of neutrino interactions with nuclei and their uncertainties in a broad kinematic range is required to distinguish signal from background and minimize systematic errors. A key ingredient of such models are the amplitudes and cross sections at the nucleon level.

In particular, a source of uncertainty arises from the nucleon axial form factor F_A. This fundamental nucleon property is a function of Q^2, defined as minus the four-momentum transferred to the nucleon squared. The axial coupling $g_A = F_A(Q^2 = 0)$ is known rather precisely from the neutron β-decay asymmetry [2],

$$g_A = 1.2723 \pm 0.0023, \tag{1}$$

although a more precise value can be obtained using recent measurements of the nucleon lifetime [3]. For the Q^2 dependence, the most common parametrization is the dipole ansatz,

$$F_A^{\text{dipole}}(Q^2) = g_A \left(1 + \frac{Q^2}{M_A^2}\right)^{-2}, \tag{2}$$

in terms of a single parameter, the so-called axial mass M_A. The dipole parametrization has been utilized to describe also the electric and magnetic form factors of the nucleon. In the Breit frame and for small momenta, this Q^2 dependence implies that the charge distribution is an exponentially decreasing function of the radial coordinate. Both the electric and the magnetic form factors of the nucleon deviate from the dipole parametrization, for a review see Ref. [4]. It seems then natural to expect similar deviations for the axial one.

Empirical information about F_A can be obtained from neutrino charged-current quasielastic (CCQE) scattering $\nu_l n \rightarrow l^- p$. Modern neutrino cross-section measurements have been performed on heavy nuclear targets (mostly ^{12}C) where the determination of F_A becomes unreliable due to the presence of not well-constrained nuclear corrections and the difficulties in isolating the CCQE channel in a model-independent way. A detailed discussion of this problem can be found, for instance, in Sec. III of Ref. [5]. A more direct and, in principle, less model-dependent determination of F_A relies on bubble-chamber data on deuterium. Global analyses of the Argonne National Laboratory (ANL) [6–8], Brookhaven National Laboratory [9,10], Fermilab [11] and CERN [12] data with updated vector from factors based on modern electron-scattering data have been performed by Bodek and collaborators. A reference value of $M_A = 1.016 \pm 0.026$ GeV with a small (2.5%) error has been obtained [13].

On the other hand, as pointed out in Refs. [14,15], anzatz (2) is not theoretically well founded. A new extraction of F_A has been recently undertaken using a functional representation of the form factor based on conformal mapping (z expansion) [16]. The function is only constrained by the analytic structure and asymptotic behavior dictated by QCD. The resulting form factor is consistent with the dipole one but with a much larger error, Fig. 7 of Ref. [16]. In particular, the axial radius,

$$r_A^2 \equiv -\frac{6}{g_A} \frac{dF_A}{dQ^2}\bigg|_{Q^2=0} \tag{3}$$

obtained is $r_A^2 = 0.46 \pm 0.22$ fm^2, which agrees with the dipole one $r_A^2 = 12/M_A^2$ but with an ~ 20 times larger error. This result might have implications for oscillation studies and calls for a new measurement of neutrino-nucleon cross sections, which is in any case desirable. The axial radius can also be extracted from muon capture by protons. A recent analysis [17] using the z expansion obtains $r_A^2 = 0.43 \pm 0.24$ fm^2, in agreement with the neutrino-scattering result.

A promising source of information about $F_A(Q^2)$ is lattice QCD. Although the experimental value of g_A has been recurrently underpredicted in lattice QCD studies, the use of improved algorithms has recently lead to consistent results [18–21]. A global analysis of the low-Q^2 and light-quark mass dependence of the results of Refs. [19–21] using baryon chiral perturbation theory has found $g_A = 1.237 \pm 0.074$ and $r_A^2 = 0.263 \pm 0.038$ [22]. The central value of r_A^2 is considerably lower than those from empirical determinations but within the (large) error bars of the z-expansion results.

The choice of a specific functional form of F_A may bias the results of the analysis. Moreover, the choice of the number of parameters within a given parametrization is a delicate question. Too few parameters may not give enough versatility. As the number of parameters increase, the χ^2 value of the fits can be reduced, but, at some point, the fit tends to reproduce statistical fluctuations of the experimental data [23]. A reduction of the model dependence of the results can be obtained within the methods of neural networks. This approach has been used to obtain nucleon parton distribution functions from deep-inelastic scattering data by the neural network parton distribution function (NNPDF) collaboration [24].

In this paper, we demonstrate that model-independent information about F_A can be obtained from a semiparametric analysis of ν-deuteron scattering data.[1] In contrast to the parametric approach in which a particular parametrization of F_A is adopted based on physics assumptions, semiparametric ones are not motivated by physics; they allow to construct a statistical model in terms of an ensemble of probability densities that are used to do statistical inference, i.e., to determine the quantities of interest and their uncertainties (see Chap. 2 of Ref. [25]). The lack of physics motivation may prevent the results from being extrapolated outside the fit region (positive Q^2 in our case). On the other hand, given the generality of the approach, the results may contain new physics beyond the underlying assumptions of a given model or be affected by theoretical mismodeling and/or deficiencies in a data set.

To perform this semiparametric analysis, we make use of feedforward neural networks,[2] a class of functions with unlimited adaptive abilities [26]. With this choice, we can eliminate any bias in the result introduced by the functional form of the fit function. Depending on the number of adaptive parameters, one can get different variants of the statistical model (fit). In this context, Bayesian statistics has proved

to be a very effective tool [27]. Its methods allow to make comparisons between different models and control the number of parameters in the fit. Bayesian methods are successfully used in different branches of physics, For instance, in hadron and nuclear physics, they have been applied to the study of the resonance content of the $p(\gamma, K^+)\Lambda$ reaction [28] and to constrain the nuclear energy-density functional from nuclear-mass measurements [29]. We consider the Bayesian framework for neural networks formulated by MacKay [30]. It has been adapted to model electric and magnetic form factors [31]. It was also used in the investigation of the two-photon exchange phenomenon in elastic electron-proton scattering [32–34]. Furthermore, this approach has proved valuable to gain insight into the proton radius puzzle and, in particular, to study the model dependence in the extraction of the proton radius from the electron-scattering data [23,35].

In the present paper, we employ the Bayesian framework for neural networks to find the most favorable $F_A(Q^2)$ based on the neutrino-deuteron CCQE scattering data measured by the ANL experiment.[3] In Sec. II, the Bayesian approach for the feedforward neural networks is introduced. Section III introduces the ANL data and theoretical framework which describes the neutrino-deuteron scattering. In Sec. IV, the numerical results are presented and discussed. In Sec. V, the summary of the study is given. Appendix A contains some details about the prior distributions in the Bayesian approach, whereas Appendix B contains the analytic form of some of the fits.

II. BAYESIAN FRAMEWORK FOR NEURAL NETWORKS

This section reviews the Bayesian approach formulated for the feedforward neural networks and its adaptation to the problem of the extraction of the nucleon axial form factor that best represents a given set of data. The proposed framework is quite general: It does not rely on physics assumptions about the functional form of $F_A(Q^2)$ and is independent of the experimental conditions from which the data originate. In this way, the present approach not only complements those based on physically motivated parametrizations, but also has the potential to disclose new physics effects as well as deficiencies in the theoretical modeling or in the data.

The general idea is the following: Given a data set, a statistical model is built. The model is characterized by a number of probability densities, which are obtained using feedforward neural networks. A detailed account of the different ingredients of the approach is given in this section. The specific application to ANL CCQE neutrino-deuteron scattering data is left for the subsequent sections.

A. Neural networks

Our aim is to obtain a statistical model which has the ability to generate $F_A(Q^2)$ values together with uncertainties. In practice, to construct such a model, a number of probability

---

[1]See Sec. II A for more detailed descriptions of parametric and semiparametric techniques.

[2]Semiparametric analyses of experimental data can also rely on other families of functions, such as polynomials or radial-basis functions [25].

[3]A global analysis including data from other experiments will be addressed in a subsequent study.

densities must be estimated. This can be achieved within three general methods [25]: (i) nonparametric, (ii) parametric, and (iii) semiparametric. In the first approach, no particular functional model is assumed, and the probabilities are determined only by the data. However, if the size of the data is large, the method requires introduction of many internal parameters. Additionally, this approach is computationally expensive. In the parametric method, a specific functional form of the model is assumed. In this case, it is relatively easy to find the optimal configuration of the model parameters. However, a particular choice of the parametrization limits the ability of the model for an accurate description of the data.[4] In this case, the uncertainties for the model prediction are either overestimated or underestimated. The semiparametric method takes the best features from both (i) and (ii) approaches. In this method, instead of a single specific functional model, a broad class of functions is considered. The optimal model is chosen among them. The neural-network approach is a realization of the semiparametric method. In particular, the feedforward neutral networks form a class of functions with unlimited adaptive abilities.

B. Multilayer perceptron

In order to model the nucleon axial form factor, a feedforward neural network in a multilayer perceptron (MLP) configuration is considered. The concept of MLP comes from neuroscience [36]. A given MLP is a nonlinear map from the input space of dimension n_i to the output space of dimension n_o,

$$\mathcal{N}: \mathbb{R}^{n_i} \mapsto \mathbb{R}^{n_o}. \tag{4}$$

The MLP map can be represented by a graph which consists of several layers of units: the input layer with n_i units, one or more hidden layers with hidden units, and the output layer which has n_o units. In the input and in every hidden layer, there is an additional bias unit. The units from the consecutive layers are all connected, but the bias unit is connected only to the following layer. As an example, the graphical representation of the MLP: $\mathcal{N}: \mathbb{R} \mapsto \mathbb{R}$ is given in Fig. 1. Every edge (connection line) in the graph represents one parameter of the function, called latter a weight.

To every unit (blue circles in Fig. 1), a real single-valued function called the activation function f is associated; its argument is the weighted sum of the activation function values received from the connected units. In the feedforward case, the ith unit in the kth layer is given in terms of the input from the $(k-1)$th layer by

$$y_{i,k} = f^{i,k} \left(\sum_{u \in \text{previous layer}} w_u^{i,k} y_{u,k-1} \right). \tag{5}$$

A graphical representation of the above function is given in Fig. 2. The weights $w_u^{i,k}$ are real parameters. Their optimal

---

[4]Fitting the axial form factor with the dipole parametrization is an example of the parametric approach.
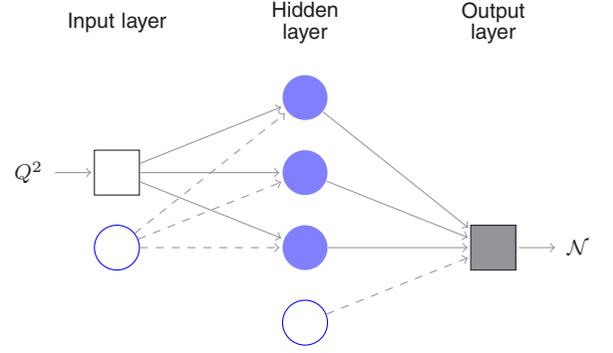

FIG. 1. Feedforward neural network in an MLP configuration $\mathcal{N}: \mathbb{R} \mapsto \mathbb{R}$. It consists of: (i) an input layer with one input unit Q^2 (open square); (ii) one hidden layer with three hidden $M = 3$ units (filled blue circles); (iii) an output layer consisting of one output unit (black square). Each line denotes a weight parameter w_j. The bias weights are denoted by dashed lines, whereas the bias units are represented by open blue circles.

values are established by the network training for which we adopt the Bayesian framework explained below.

Note that for the bias unit $f^{i,k}(x) = 1$. Furthermore, it is assumed that in the output layer the activation functions are linear $f(x) = x$. In order to simplify and speed up the performance of the numerical analyses, MLPs with only one hidden layer of units are considered. In Fig. 1, there is an example of such an MLP with $M = 3$ hidden units.

Let us introduce the MLP $\mathcal{N}_M: \mathbb{R} \mapsto \mathbb{R}$ with a single hidden layer and M units, which has the following functional form:

$$\mathcal{N}_M(Q^2; \{w_j\}) = \sum_{n=1}^{M} w_{2M+n} f\left(w_n \frac{Q^2}{Q_0^2} + w_{M+n} \right) + w_{3M+1},$$
$$\tag{6}$$

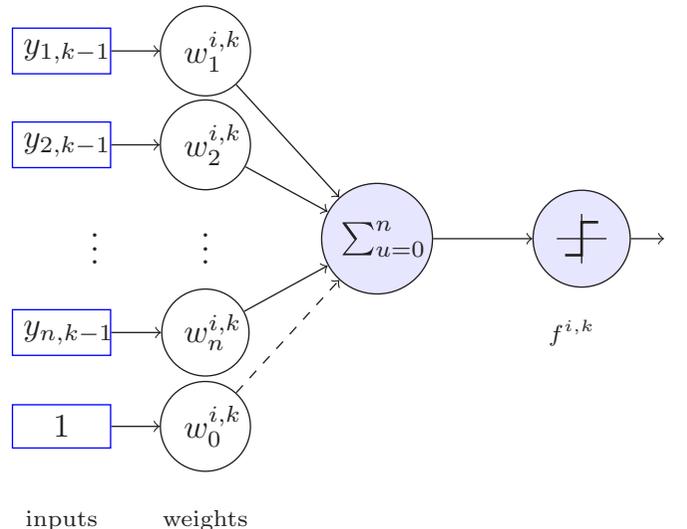where $Q_0^2 \equiv 1$ GeV^2. This function depends on $W = 3M + 1$ weights and Q^2.

FIG. 2. The ith unit in the kth layer [Eq. (5)].

It has been proved and demonstrated (Cybenko theorem) [26,37–42] that, if M is sufficiently large, feedforward neural networks with sigmoidal and linear activation functions in the (single) hidden and output layers, respectively, form a dense subset in the set of continuous functions. This implies that a map of the form (6) can approximate arbitrarily well any continuous function and its derivative. As required by the theorem, in our numerical analysis, the activation functions in the hidden layer are given by sigmoids,

$$f(x) = \frac{1}{1 + \exp(-x)}. \tag{7}$$

C. Axial form factor

We seek for a model-independent parametrization of F_A that best represents the available data without any input from theory. It should be quite general but, nonetheless, constrained by the following general properties:

(I) $F_A(Q^2)$ is assumed to be a continuous function of Q^2 in its validity domain;

(II) the domain of F_A is restricted to the $Q^2 \in (0, 3)$ GeV^2 where the ANL data are present;

(III) $F_A(Q^2 = 0)$ is constrained by the g_A experimental value Eq. (1);

(IV) as $F_A(Q^2)$ is bounded, there must be a $C > 1$: $F_A(Q^2) < C F_A^{\text{dipole}}(Q^2)$ in the whole Q^2 interval of $(0, 3)$ GeV^2.

The feedforward neural network of Eq. (6) can fulfill these properties; for a more detailed discussion see Appendix A. In order to speed up the numerical computations, we rescale the output of the MLP (6) by normalizing it to the dipole ansatz. As the result, the axial form factor is represented by

$$F_A(Q^2) = F_A^{\text{dipole}}(Q^2) \times \mathcal{N}_M(Q^2; \{w_i\}), \tag{8}$$

where F_A^{dipole} is given in Eq. (2) with $M_A = 1$ GeV. In this way, the neural-network response gives the deviation of the axial form factor from the dipole parametrization. The value of g_A is allowed to change within the Particle Data Group's (PDG) uncertainty Eq. (1).

D. Bayesian approach and Occam's razor

As described above, a MLP is a nonlinear map defined by some number of the adaptive parameters. The increase in the number of hidden units improves MLP's ability to reproduce the data. However, when the number of units (weights) is too large, the model tends to overfit the data, and it reproduces the statistical noise. As a result, its predictive power is lost. On the other hand, if the network is too small, then the data are underfitted. This competition between two extreme cases is know in statistics as the bias-variance trade-off [43]. Certainly, the optimal model is a compromise between both extreme situations.

Bayesian statistics provides methods to face the bias-variance trade-off dilemma. Indeed, the Bayesian approach naturally embodies Occam's razor [27,30,44,45], i.e., complex models, defined by a large number of parameters, are naturally penalized, whereas simple fits with a small number

of parameters are preferred. Moreover, the Bayesian approach allows one to make comparisons between different statistical descriptions of the data and to indicate the model which is favored by the measurements. An example of such analysis can be found in Ref. [23] where a large number of different fits of the electric and magnetic form factors were obtained from electron-proton scattering data. For each model, the value of the proton radius r_p^E has been calculated. It turned out that, depending on the model, r_p^E ranges from 0.8 to 1.0 fm. A considerable fraction of the results agreed with the muonic-atom measurement $r_p^E = 0.84\,184(67)$ fm [46], but the Bayesian algorithm preferred a model with $r_p^E = 0.899 \pm 0.003$ fm, which is in contradiction with the muonic-atom result but in agreement with some other non-Bayesian ep scattering data analysis. A critical review of various approaches to proton radius extraction can be found in Ref. [47] and references therein.

E. Bayesian framework for MLP

1. General idea

We adopt the Bayesian framework for the feedforward neural network formulated by MacKay [48,49]. The main concepts of the approach are briefly reviewed below.

Let us consider the set of neural networks,

$$\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_M, \tag{9}$$

where M denotes the number of units in the hidden layer. To each of the models $\mathcal{N}$, one associates a prior probability denoted $\mathcal{P}(\mathcal{N})$. Our task is to obtain two posterior conditional probabilities: $\mathcal{P}(\mathcal{N}|\mathcal{D})$—the probability of the model $\mathcal{N}$ given a data set $\mathcal{D}$ and $\mathcal{P}(\rho|\mathcal{D}, \mathcal{N})$—probability distribution of the model parameters given $\mathcal{D}$ and model assumptions; ρ denotes the set of model parameters, which encompasses the neural network weights $\rho = [\{w_j\}, \ldots]$. The first probability density allows us to choose among many network types the one which is favorable by the data, whereas the second one is necessary to make model predictions.

If one assumes, at the beginning of the analysis, that all MLP configurations are equally suited for describing the data, then the following relations between prior probabilities hold

$$\mathcal{P}(\mathcal{N}_1) = \mathcal{P}(\mathcal{N}_2) = \cdots = \mathcal{P}(\mathcal{N}_M). \tag{10}$$

Then, in order to classify the models, it is sufficient to compute the so-called evidence $\mathcal{P}(\mathcal{D}|\mathcal{N})$. Indeed, from Bayes' theorem, one gets

$$\mathcal{P}(\mathcal{N}|\mathcal{D}) = \frac{\mathcal{P}(\mathcal{D}|\mathcal{N})\mathcal{P}(\mathcal{N})}{\mathcal{P}(\mathcal{D})} \sim \mathcal{P}(\mathcal{D}|\mathcal{N}), \tag{11}$$

where $\mathcal{P}(\mathcal{D})$ is the normalization constant.

On the other hand, the posterior probability for the weights of a given MLP reads

$$\mathcal{P}(\rho|\mathcal{D}, \mathcal{N}) = \frac{\mathcal{P}(\mathcal{D}|\rho, \mathcal{N})\mathcal{P}(\rho|\mathcal{N})}{\mathcal{P}(\mathcal{D}|\mathcal{N})}, \tag{12}$$

where $\mathcal{P}(\mathcal{D}|\rho, \mathcal{N})$ is the likelihood whereas the density $\mathcal{P}(\rho|\mathcal{N})$ is the prior describing the initial assumptions about the parameters. By integrating both sides of Eq. (12), one gets

the evidence for the model,

$$\mathcal{P}(\mathcal{D}|\mathcal{N}) = \int d\rho \, \mathcal{P}(\mathcal{D}|\rho, \mathcal{N})\mathcal{P}(\rho|\mathcal{N}). \qquad (13)$$

2. Likelihood, prior, and posterior densities

In order to calculate the posterior (12), we assume that the likelihood is given in terms of the χ^2 function,

$$\mathcal{P}(\mathcal{D}|\{w_j\}, \mathcal{N}) = \frac{1}{N_L} \exp(-\chi^2), \qquad (14)$$

where N_L is the normalization constant. The χ^2 function for the present paper is defined in Sec. III B [see Eq. (25)].

It is also assumed that the initial values of the weights are Gaussian distributed (the arguments supporting this choice are collected in Appendix A)

$$\mathcal{P}(\{w_j\}, \mathcal{N}) = \left(\frac{\alpha}{2\pi}\right)^{W/2} \exp[-\alpha \, E_w(\{w_j\})], \qquad (15)$$

where α is a hyperparameter (regularizer) introduced to deal with the overfitting problem and

$$E_w(\{w_j\}) = \frac{1}{2} \sum_{i=1}^{W} w_i^2. \qquad (16)$$

The regularizer α plays a crucial role in model optimization and should be properly determined. Indeed, if α is large, then the term (16) dominates in the posterior Eq. (12), so it is very likely that the model underfits the data. On the contrary, if α is too small, the likelihood dominates, and the model tends to overfit the data. Note that α is another parameter of the model, hence, $\rho = \{\{w_j\}, \alpha\}$.

In principle, to get the evidence $\mathcal{P}(\mathcal{D}|\mathcal{N})$ on which model discrimination is based, the integration in Eq. (13) over the whole space of parameters ρ should be performed. This is, however, numerically difficult to perform. Therefore, in our analysis, we consider another method, the so-called evidence approximation [48,49].

3. Evidence approximation

In the adopted approach, it is assumed that the posterior distributions have a Gaussian shape. Hence, to get the necessary information about (12), it is enough to obtain the configuration of the parameters ρ_{MP} at which the posterior distribution is at its maximum and the covariance matrix for the model. The latter is necessary to provide the uncertainties for the model predictions.

In this approach, which corresponds to the type-II maximum likelihood method of conventional statistics [50], the optimal value ρ_{MP} is established during the training of the network. In this process, the *a priori* unknown α parameter is iteratively changed {see Eqs. (3.18) and (3.19) of Ref. [31]}, starting from small value $\alpha_0 = 0.001$, which leads to a posterior covering a large region in the parameter space. The iteration procedure is convergent, and the result has a negligible dependence on the initial α value. More details about the algorithm implementation can be found in Sec. 3.1 of Ref. [31] and in Sec. III. C of Ref. [33]. Note that the optimal
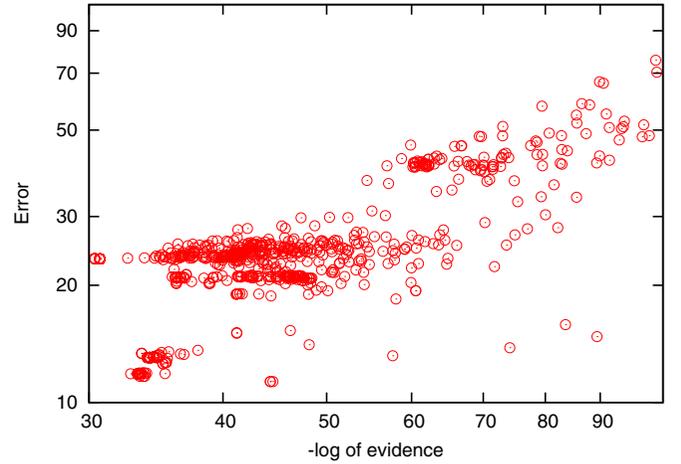

FIG. 3. The error Eq. (21) as a function of the logarithm of the evidence Eqs. (17) and (18). Each point denotes the result obtained for one MLP fit to the BIN1 data, including deuteron corrections (see Sec. III for details). The fits with a logarithm of the evidence smaller than -100 are not shown in the figure.

configuration of the model parameters $\rho_{MP} = \{\{w_j\}_{MP}, \alpha_{MP}\}$ is close to the configuration for which χ^2 is at the minimum.

Within the present approximation, the evidence for a given model is cast in an analytical form. Namely,

$$\ln \mathcal{P}(\mathcal{D}|\mathcal{N}) \approx -\chi^2 - \alpha_{MP} E_w(\{w_j\}_{MP}) \qquad (17)$$

$$-\frac{\ln|A|}{2} + \frac{W}{2} \ln \alpha_{MP} - \frac{1}{2} \ln \frac{\gamma(\alpha_{MP})}{2}$$

$$+ M \ln(2) + \ln(M!). \qquad (18)$$

In the above expression, normalization factors common to all models are omitted; $|A|$ denotes the determinant of the Hessian matrix,

$$A_{ij} = \nabla_i \nabla_j \chi^2|_{\{w_k\}=\{w_k\}_{MP}} + \delta_{ij}\alpha_{MP}. \qquad (19)$$

The parameter γ is given by

$$\gamma(\alpha) = \sum_{i=1}^{W} \frac{\lambda_i}{\alpha + \lambda_i}; \qquad (20)$$

$\gamma(\alpha_{MP})$ measures the effective number of weights, whose values are controlled by the data [25]. The λ_i's are eigenvalues of the matrix $\nabla_i \nabla_j \chi^2$.

The evidence contains two contributions: Occam's factor [(18) plus the $\alpha_{MP} E_w$ term of (17)] which is large for models with many parameters and the misfit [χ^2 term in Eq. (17)], which could be large if the model is too simple. Therefore, the model which maximizes the evidence is the one which solves the bias-variance dilemma. As an illustration from the present paper (details can be found in the following section), in Fig. 3 we plot the values of the error,

$$\mathcal{E} = \chi^2 + \alpha E_w, \qquad (21)$$

and the evidence for MLP fits. The best model with the highest evidence is not the one which has the smallest value of the error function $\mathcal{E}$, in variance with more conventional

approaches based on the minimization of the χ^2 per degree of freedom.

III. ANALYSIS OF ANL NEUTRINO-DEUTERON-SCATTERING DATA

A. Theoretical framework

The neutrino-induced CCQE,

$$\nu_\mu(k) + n(p) \to \mu^-(k') + p(p'), \tag{22}$$

differential cross section, in terms of the standard Mandelstam variables $s = (k + p)^2$, $u = (p - k')^2$, and $t = (k - k')^2 = -Q^2$, can be cast as [51]

$$\frac{d\sigma}{dQ^2} = \frac{G_F^2 m_N^2}{8\pi E_\nu^2}\left[A(Q^2) + B(Q^2)\frac{(s-u)}{m_N^2} + C(Q^2)\frac{(s-u)^2}{m_N^4}\right], \tag{23}$$

where A, B, and C are quadratic functions of the vector $[F_{1,2}^V(Q^2)]$ and axial $[F_{A,P}(Q^2)]$ form factors {see Eqs. (3.18) of Ref. [51]}; G_F is the Fermi constant, m_N is the nucleon mass, and E_ν is the neutrino energy in the laboratory frame. Isospin symmetry allows to relate $F_{1,2}^V$ to the corresponding electromagnetic proton and neutron form factors, which are extracted from electron-scattering data. In the present paper, we have taken these electromagnetic form factors from Refs. [52,53]. With this simple choice, we disregard deviations from the dipole shape because the accuracy of the neutrino-deuteron data is insufficient to be sensitive to them, particularly, at the rather low $Q^2 \lesssim 1$ GeV^2 probed in the ANL experiment. Finally, the partial conservation of the axial current (PCAC) and the pion-pole dominance of the pseudoscalar form factor F_P allow to express it in terms of F_A.

Deuterium-filled bubble-chamber experiments actually measured $\nu_\mu + d \to \mu^- + p + p$. The cross section for this process differs from the one on free neutrons due to the momentum distribution of the neutron in the nucleus, Pauli principle, final-state interactions, and meson-exchange currents. In the literature, it has been commonly assumed that Eq. (23) can be corrected for these effects by a multiplicative function of Q^2 alone $R(Q^2)$ and such that $R \to 1$ at large Q^2. For the present paper, we adopt $R(Q^2)$ from the calculation of Ref. [54] (solid line in Fig. 4).

B. χ^2 function for the ANL experiment

In the ANL experiment, the interactions of muon neutrinos in a 12-ft bubble chamber filled with liquid deuterium were studied [6–8,55]. The neutrino flux peaked at $E_\nu \sim 0.5$ GeV and has fallen by an order of magnitude at $E_\nu = 2$ GeV [7,8]. For the statistical analysis, we consider the Q^2 distribution of CCQE events. Some of the originally published bins were combined together to have a number of events larger than five in every bin. Therefore, the number of bins is $n_{\rm ANL} = 25$ where bins from 1 to 23 have a width of 0.05 GeV^2, whereas bins 24 and 25 have widths of 0.65 GeV^2. The total number of measured two- and three-prong events adds to $N_{\rm ANL} = 1792$ [8]. One-prong events were not included in the ANL selection. To account for their loss, the region of $Q^2 = 0.05$ GeV^2 was excluded.
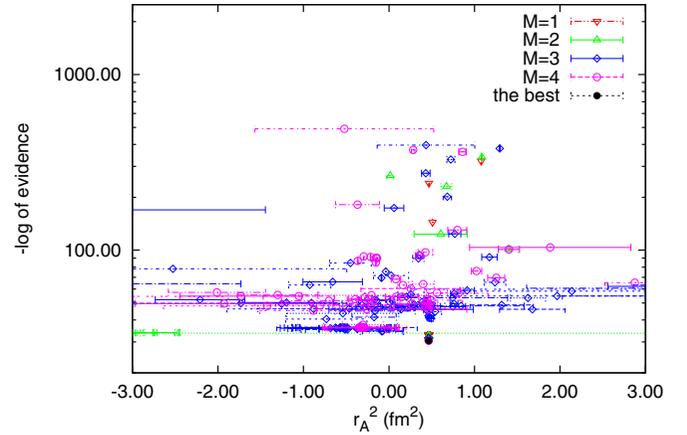

FIG. 4. The dependence of r_A^2, defined in Eq. (3), on the logarithm of the evidence (17) and (18). Results for the MLP fits to BIN1 data (without the deuteron correction). The MLPs consist of $M = 1$–4 hidden units.

The predicted number of events in each bin is calculated similarly as in Ref. [16],

$$N_i^{\rm th} = p\int_0^\infty dE_\nu \frac{\frac{d\sigma}{dQ^2}(E_\nu, F_A, Q_i^2)}{\sigma(E_\nu, F_A)}\frac{dN}{dE_\nu}, \tag{24}$$

where $p(dN/dE_\nu)/\sigma(E_\nu, F_A)$ is the neutrino energy flux, given in terms of the experimental energy distribution of observed events dN/dE_ν taken from Ref. [55].

As stated in the previous section, the likelihood [Eq. (14)] is built using the χ^2 function, which we cast as

$$\chi^2 = \chi^2_{\rm ANL} + \chi^2_{g_A}, \tag{25}$$

where $\chi^2_{g_A}$ is introduced to constrain the value of the axial form factor at $Q^2 = 0$,

$$\chi^2_{g_A} = \left(\frac{F_A(0) - g_A}{\Delta g_A}\right)^2; \tag{26}$$

g_A and Δg_A are fixed by the present PDG central value and its uncertainty, respectively, Eq. (1). For $\chi^2_{\rm ANL}$, we take

$$\chi^2_{\rm ANL} = \sum_{i=k}^{n_{\rm ANL}}\frac{(N_i - N_i^{\rm th})^2}{N_i} + \left(\frac{1-p}{\Delta p}\right)^2, \tag{27}$$

where N_i denotes the number of events in the bin. The last term takes into account the systematic uncertainty in the total number of events [56] inherited from the flux-normalization uncertainty. Similarly, as in the analysis of single-pion production data [57], it is assumed that $\Delta p = 0.20$.[5] At the beginning of the analysis, $p = 1$ is set. Then, during the training of the network, p is iteratively updated. This algorithm is described in Ref. [32].

It is known that the low-Q^2 data are characterized by a lower efficiency (see, for instance, Fig. 1 of Ref. [8]). Moreover, in this kinematic domain, deuteron structure corrections

---

[5]This is a more conservative value of the flux-normalization uncertainty than the ANL estimate of 15% [8].

must be carefully discussed. In order to study this problem, we consider three variants of the ANL data:

(i) $\chi^2_{\text{ANL}} \to \chi^2_{\text{BIN0}}$: All ANL bins included;
(ii) $\chi^2_{\text{ANL}} \to \chi^2_{\text{BIN}k}$: where $k = 1$ or $k = 2$: ANL bins without the first k bins.

Additionally, for each data set, we consider the cross-section model both with and without $[R(Q^2) \equiv 1]$ deuteron corrections.

C. Numerical algorithm

We consider a MLP with $M = 1$–4 hidden units in a single hidden layer. For $M > 4$, the number of parameters in the fit starts to be comparable with number of bins. The numerical algorithm for getting the optimal fit is summarized by the following list of steps:

(1) Consider a MLP with a fixed number of hidden units $M = 1$;
(2) using the Bayesian learning algorithm ([33]), perform the network training and find the optimal values for the weights and the regularizer α;
   (a) set the initial value of $\alpha \equiv \alpha_0 = 0.001$;
   (b) initialize randomly the values of the weights;
   (c) perform training until the maximum of the posterior is reached; at each iteration step, update the values of weights and α.
(3) Calculate the evidence for each of the obtained MLP fits;
(4) repeat steps (1)–(3) for various initial configurations of $\{w_j\}$;
(5) among all registered fits, choose the best one according to the evidence;
(6) repeat steps (1)–(5) for $M = 2$–4;
(7) among the best fits obtained for $\mathcal{N}_{1-4}$ MLPs, choose the model with the highest evidence.

The optimal configuration of parameters is obtained using the Levenberg-Marquardt algorithm [58,59].

IV. NUMERICAL RESULTS

The analysis of the BIN0, BIN1, and BIN2 data sets has been independently performed. For each set, both cross-section models with and without deuteron corrections have been studied. For the default analyses, Δg_A has been taken from the PDG as in Eq. (1), but the impact of a larger uncertainty $\Delta g_A / g_A = 10\%$ has been investigated and is discussed below. We have also performed analyses with normalization uncertainties smaller ($\Delta p = 0.10$) and larger ($\Delta p = 0.30$) than the default $\Delta p = 0.20$, but it turned out that decreasing or increasing Δp does not significantly affect the final results. All in all, about 17 000 fits have been collected. Among them, for each type of analysis, the best model has been chosen according to the objective Bayesian criterion described in Sec. II.

In order to compare quantitatively different analyses, one needs to take into account the relative data normalization $\mathcal{P}(\mathcal{D})$. This density is not evaluated within the adopted
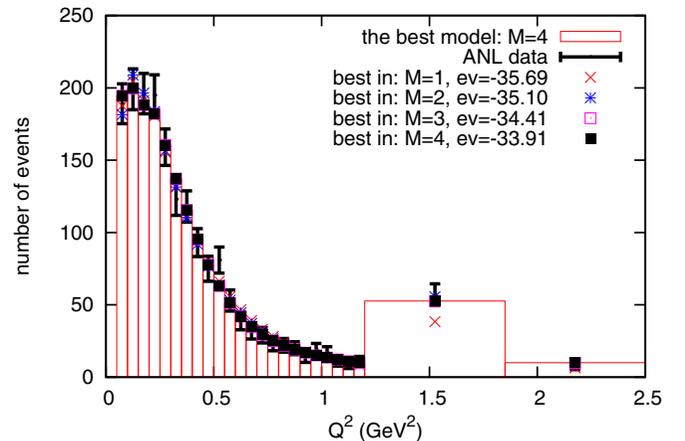

FIG. 5. Distribution of the ANL number of events and the best fits obtained for MLPs with $M = 1$–4 hidden units. The figure shows the results of the analysis of BIN0 data (deuteron corrections included). For each fit, the value of the logarithm of evidence (ev) is given.

approach. Hence, we cannot quantitatively compare the results of, e.g., BIN0 and BIN1 analyses. Nonetheless, for a given data set, quantitative comparisons between the results obtained within with the two versions of the cross-section model can be made.

As described in Sec. III C for each type of analysis (data set plus cross-section model), to find the optimal fit, MLPs with: $M = 1$–4, hidden units have been trained. The best model within each MLP type is the one with the maximal value of the evidence Eqs. (17) and (18).

In order to illustrate the performance of the training algorithm, in Fig. 4 we present the dependence of the resulting axial-radius squared (r_A^2) values on the evidence for the BIN1 data set. The best fit with the highest evidence, obtained with $M = 4$, gives $r_A^2 \approx 0.464$ fm^2.

Note that all the best models within each MLP type reproduce well the ANL data. This is illustrated in Fig. 5, which presents the distribution of the ANL events and the best fits.

Our main results, i.e., the best fits to BIN0, BIN1, and BIN2 data for the model with and without the deuteron correction, and Δg_A from Eq. (1) are summarized in Table I.

TABLE I. The best MLP fits, obtained for the analysis of the BIN0, BIN1, and BIN2 data with and without deuteron corrections; Δg_A is taken from Eq. (1).

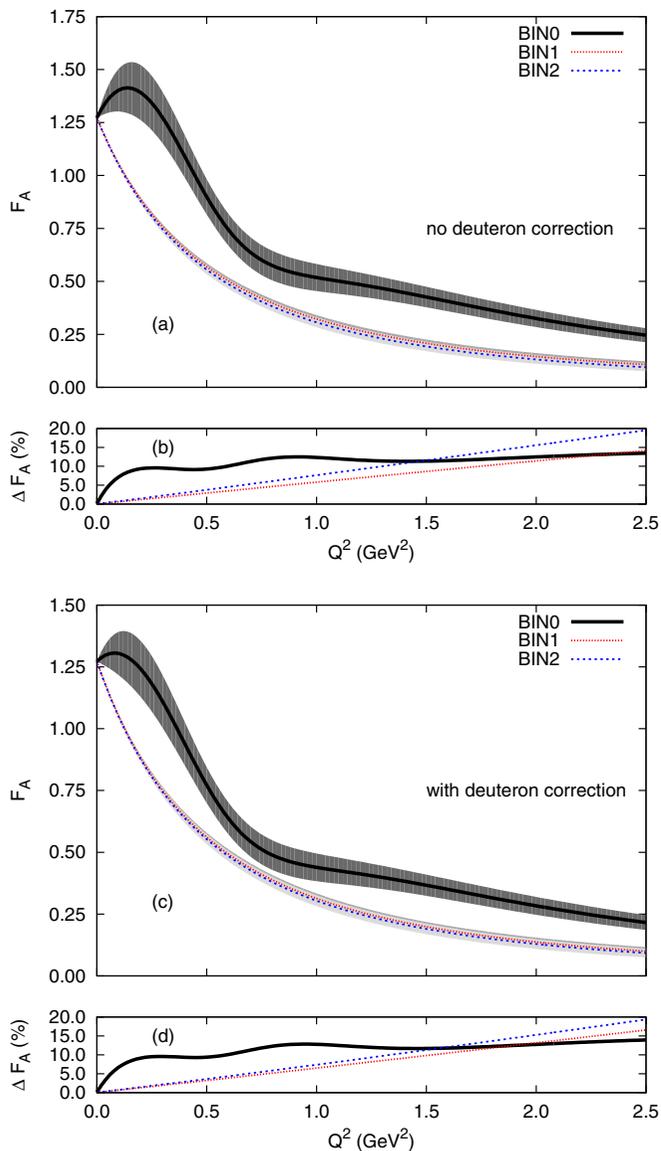| Deuteron	M	$\ln \mathcal{P}(\mathcal{D}, \mathcal{N})$	χ^2	p	$\mathcal{E}$	r_A^2 (fm^2)
			BIN0			
No	4	-34.72	11.97	1.1	14.14	-0.394 ± 0.278
Yes	4	-33.91	11.73	1.08	13.95	-0.161 ± 0.240
			BIN1			
No	4	-30.57	24.84	1.16	25.41	0.464 ± 0.014
Yes	3	-29.6	22.90	1.12	23.43	0.471 ± 0.015
			BIN2			
No	2	-30.15	22.62	1.18	23.16	0.476 ± 0.017
Yes	4	-27.67	21.94	1.13	22.62	0.478 ± 0.017

FIG. 6. Best fits of the axial form factor obtained from the analysis of the BIN0, BIN1, and BIN2 data sets. The top (bottom) panel presents the results obtained without (with) deuteron corrections. The shaded areas denote 1σ uncertainties. Additionally, the relative uncertainty $\Delta F_A/F_A$ is plotted.

The corresponding $F_A(Q^2)$ functions with their error bands are shown in Fig. 6.

Both fits for the BIN0 data set, which contains all the data from the original ANL measurement with and without deuteron corrections show a Q^2 behavior of F_A with a rapid increase followed by a decrease after a local maximum. As a result, r_A^2 has a negative sign,[6] which is at odds with all available determinations. We also observe that the Q^2 dependence of the form factor disagrees with the one obtained from the same data set using the z expansion {the coefficients from the ANL fit are given in Eq. (19) of Ref. [16]}. We have obtained

the z-expansion coefficients for the BIN0 best fit, finding that their values grow with the expansion order to values that are too large compared to phenomenological expectations [14]. This is an indication that the fit that best represents BIN0 data is inconsistent with the QCD assumptions implicit in the z expansion.

The height of the F_A maximum is reduced once the deuteron correction is included in the analysis, and it disappears when the first bin is removed from the ANL data (BIN1 data set).[7] Hence, the presence of the local maximum of F_A appears to be caused by low-Q^2 effects. Furthermore, the coefficients of the z expansion for the fits to BIN1 and BIN2 data sets are fully consistent with the expectations from QCD.

There are several possible sources of this unexpected behavior of the fits to the BIN0 set, namely, (i) an improper description of the nuclear corrections; (ii) a low quality of the measurements at low-Q^2 due to low and not well-understood efficiency; (iii) constraints coming from the uncertainty of g_A; (iv) because of the lack of very low-Q^2 data, the actual value of r_A^2 might not be properly estimated: For instance, if F_A has first a local minimum and then a local maximum.[8] In the later scenario, the ANL data (and the available bubble-chamber data, in general) are not precise enough to reveal this behavior.

In the low-Q^2 kinematic domain, deuteron effects are sizable and may play a crucial role. On the other hand, the inclusion of deuteron corrections in the analysis of the BIN1 and BIN2 data sets has a minor impact on the functional dependence of the final results, i.e., there is small difference between $F_A(Q^2)$ obtained with and without deuteron corrections as can be seen in Fig. 7. It is also interesting to highlight that the inclusion of the deuteron-structure corrections in the cross-section model increases the value of the evidence for each type of the analysis, see Table I.

In Fig. 8, we plot values of r_A^2 against the evidence. It is clearly seen that the fits including deuteron corrections are favored by the ANL data. The impact of the sensitivity of the results on the deuteron structure revealed in the present paper calls for a more accurate account of this ingredient of the cross-section models, beyond the $R(Q^2)$ function from Ref. [54] employed so far. Recent studies of CC νd scattering in the QE regime (without pions in the final state) include the nonrelativistic calculation of the inclusive cross section, incorporating two-body amplitudes, of Ref. [60]. For the kinematics of the ANL and other bubble-chamber experiments, it is important to employ a relativistic framework as in Ref. [61]. Furthermore, the consideration of the semi-inclusive rather than the inclusive cross section will allow taking into account the detection threshold for outgoing protons, which, in the ANL case, is 100 MeV [7]. One should nonetheless bear in mind that even with the best model for the deuteron there is no guarantee that the low-Q^2 region is successfully described

---

[6]Although the large uncertainty does not exclude positive values.

---

[7]It is worth mentioning that fits with a negative slope of F_A at low Q^2, resembling the best result for BIN0 data, have also been obtained in this case, but they are not preferred by the Bayesian algorithm.

[8]The magnetic form factors of the nucleon at very low-Q^2 (about 0.01 GeV^2) when normalized to a dipole have an oscillatory Q^2 dependence [31].
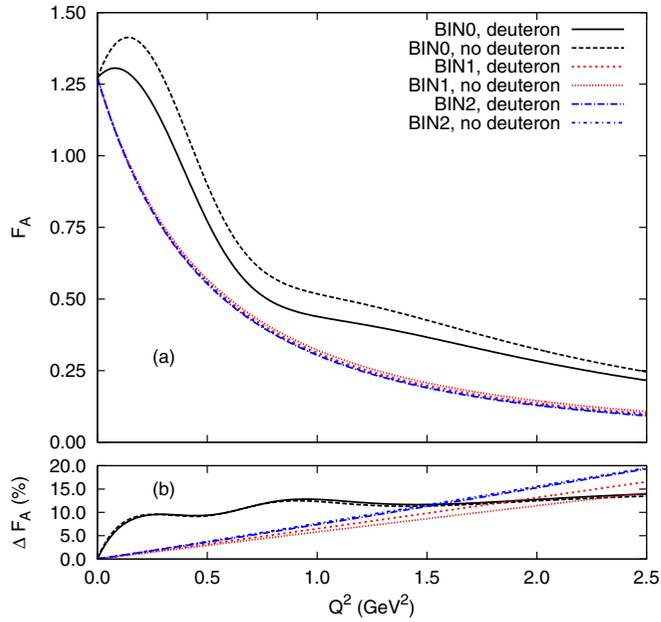
FIG. 7. Impact of the deuteron corrections on the axial form factor fits. Results of the best fits to the BIN0, BIN1, and BIN2 data sets with and without the deuteron correction together with relative uncertainties. All curves for the BIN1 and BIN2 cases nearly overlap.

because of the difficulties in the measurement and with efficiency estimates at this kinematics.

The impact of Δg_A on the results can be easily investigated. Indeed, if one increases the Δg_A uncertainty from $\Delta g_A/g_A \approx 0.1\%$ as in Eq. (1) to 10%, then the local maximum of F_A disappears. However, the fit uncertainty rapidly grows from $\Delta F_A/F_A$ lower than 0.01% to $\Delta F_A/F_A \approx 7\%$ at $Q^2 = 0$. This analysis is shown in Fig. 9.

In order to compare the Bayesian neural-network results with the traditional approach, we have performed a
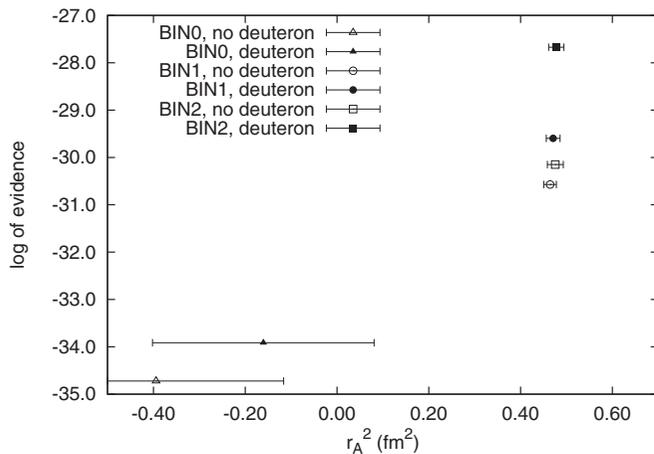

FIG. 8. Dependence of r_A^2 on the logarithm of evidence. Open and full triangles denote (the best) fits to the BIN0 data without and with the deuteron corrections, respectively. Analogously, the fits to the BIN1 and BIN2 data are denoted by circles and squares, respectively.
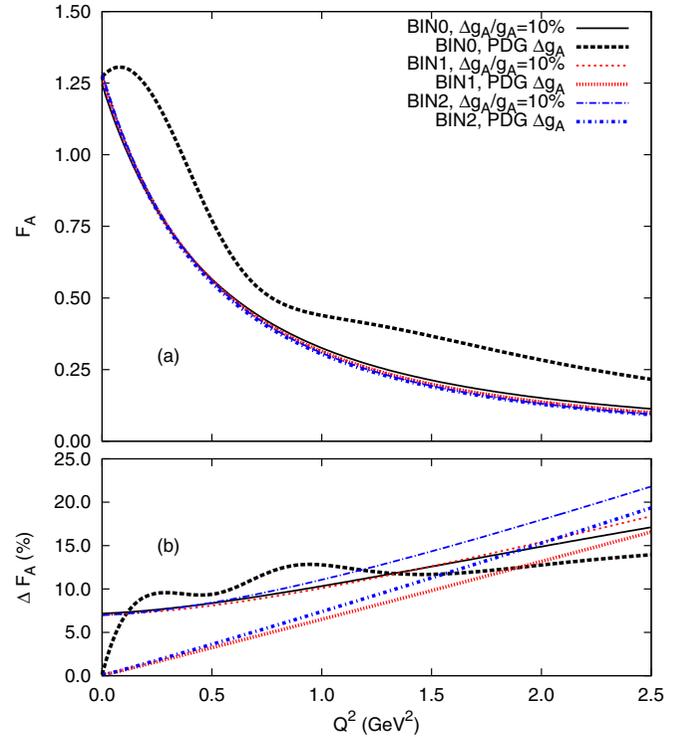

FIG. 9. Impact of the Δg_A uncertainty on the extraction of the axial form factor. The best fits to BIN0, BIN1, and BIN2 data (with deuteron correction). The thin lines denote results obtained assuming a 10% uncertainty for g_A, whereas the thick lines denote fits with the PDG uncertainty.

conventional analysis of the ANL data assuming the dipole parametrization for the axial form factor Eq. (2). The best fit minimizes the $\chi^2_{\rm ANL}$ function [Eq. (27)].[9] These results are summarized in Table II, whereas the comparison between dipole fits and neural-network analyses are displayed in Fig. 10.

Let us stress that the r_A^2 and normalization parameter p for the fits of BIN0 data are comparable to the z-expansion

---

[9]For these analyses, the MINUIT package of ROOT has been utilized.

TABLE II. Fits of the dipole axial form factor (2) to the BIN0, BIN1, and BIN2 data sets.

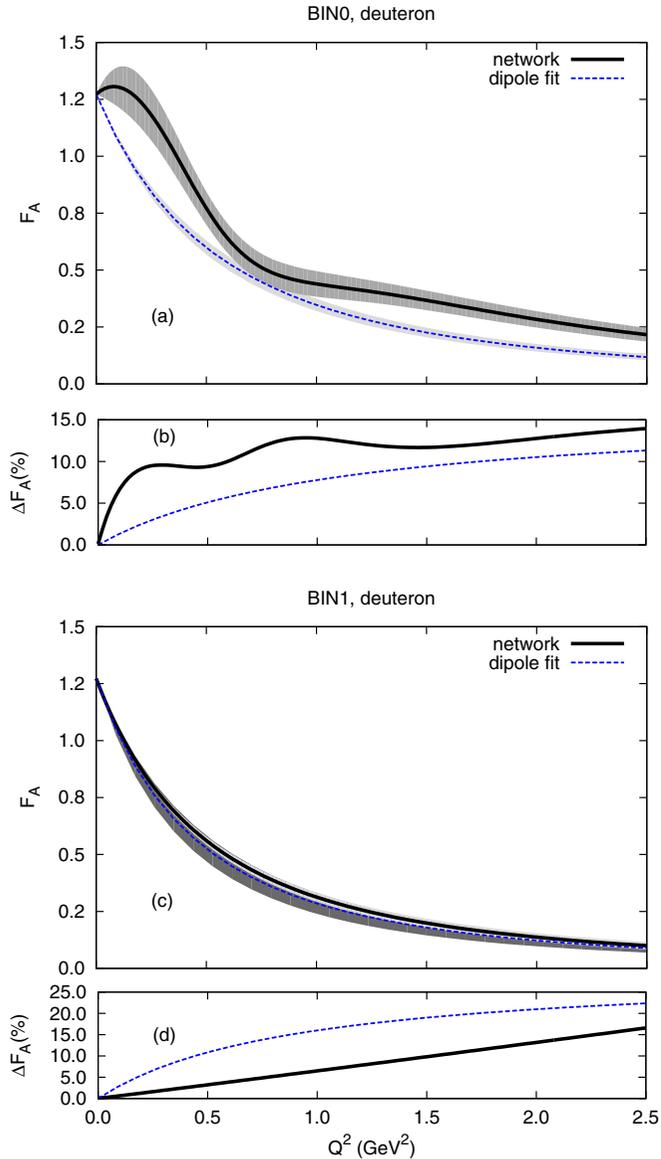|	χ^2 analyses			
	χ^2	p	M_A (MeV)	r_A^2 (fm^2)
		BIN0		
No deuteron	33.3	1.12 ± 0.03	1110 ± 60	0.38 ± 0.04
Deuteron	28.0	1.09 ± 0.03	1050 ± 60	0.43 ± 0.05
		BIN0		
No deuteron	24.4	1.17 ± 0.03	1000 ± 70	1.47 ± 0.07
Deuteron	22.3	1.13 ± 0.03	950 ± 70	0.52 ± 0.08
		BIN2		
No deuteron	20.8	1.22 ± 0.05	890 ± 100	0.59 ± 0.13
Deuteron	19.8	1.18 ± 0.05	850 ± 110	0.65 ± 0.16

FIG. 10. Comparison of the dipole with the neural network fits to the BIN0 and BIN1 data, the deuteron corrections included. The shaded areas denote 1σ uncertainties of F_A.

results [16], even though in, the latter case, a different error function was utilized. Certainly, with a dipole fit to BIN0 data one cannot obtain the local maximum of F_A at low Q^2. On the other hand, the dipole fits to the BIN1 and the BIN2 data have very similar functional Q^2 dependence as the best MLP fits. For these data sets, the evidence, which contains Occam's factor penalizing overfitting parametrizations with large error bars, establishes the preference for a rather structureless neural network that departs very little from the normalization values, given by Eq. (2) with $M_A = 1$ GeV, and has small errors. Furthermore, the uncertainties in the neural-network fits are systematically smaller than in the dipole χ^2 ones.

V. SUMMARY

The first Bayesian neural-network analysis of the neutrino-deuteron scattering data has been performed. The reported study has been oriented to the extraction of the axial form factor from the ANL CCQE data, searching for deviations from the dipole form. With the full ANL data set, the analysis leads to an axial form factor which has a positive slope at $Q^2 = 0$ and a local maximum at low Q^2. The inclusion of the deuteron correction reduces the peak in F_A. Only after removing the lowest available Q^2 region ($0.05 < Q^2 < 0.10$ GeV^2) from the data, a value of the axial radius consistent with available determinations could be obtained. This suggests that corrections from the deuteron structure play a crucial role at low Q^2 but it could also be the case that the experimental errors in this kinematic region were underestimated. Analyses without the low-Q^2 data do not show any significant deviation from previous determinations. Furthermore, our neural-network fits are characterized by smaller uncertainties than the dipole ones.

New more precise measurements of the neutrino cross sections on hydrogen and deuterium are needed to unravel the axial structure of the nucleon. Techniques, such as the one applied in the present paper, will prove valuable in such a scenario.

APPENDIX A: PRIOR DISTRIBUTIONS

The prior distribution of the weights, Eqs. (15) and (16), is justified by the following properties of the adopted feed-forward neural network in the MLP configuration and of the problem under study

(P1) internal symmetry: The exchange of any two units in the hidden layer does not change the functional form of the network and its output values;

(P2) the sigmoid activation function $f(x)$ Eq. (7) saturates and can be effectively assumed to be constant outside the interval $-a \leqslant x \leqslant a$ with $a \sim 10$;

(P3) $f(-x) = 1 - f(x)$;

(P4) the ANL data are concentrated in the region $Q^2 \in (0, 3)$ GeV^2—constraint (II) in Sec. II C;

(P5) $F_A(Q^2)/F_A^{\text{dipole}}(Q^2) < C$, where $C > 1$—constraint (IV) in Sec. II C.

Properties (P4) and (P2) constrain the weights $w_1, \ldots, w_{2M}$ in function (6). Indeed, from (P2), one sees that for a full performance of the activation function $f(x)$ in Eq. (6) it is enough to have $x \in (-a, a)$. Let us consider then the function $f(w_i Q^2 + w_{M+i})$, $i = 1, \ldots, M$, which is one of the elements of the sum (6). To efficiently cover all possible outputs, achieving a good performance of the network, it is enough that the arguments of $f(\cdots)$ belong to $(-a, a)$ for all values of Q^2 under consideration. Then, one gets limits for the weights, namely, $|w_{M+i}| < a$[10] from which one also gets the constraint $|w_i| < 2/3a < a$ for $Q^2 \in (0, 3)$ GeV^2.

Note that both negative and positive values of weights w_i and w_{M+i} are equally possible according to (P3). Therefore, without losing generality, the prior density should be symmetric in weights $w_{1 \text{ to } 2M}$ and cover the hypercube $(-a, a)^{2M}$.

The limits for the weights in the linear layer are less obvious. Property (P5) provides a constraint on the weights in the linear output layer w_{2M+i}, $i = 1, \ldots, M + 1$, namely,

$$\left| \sum_{i=1}^{M} w_{2M+i} f(\cdots) + w_{3M+1} \right| < C. \tag{A1}$$

The main role of the weights in the linear layer is to control the range of the neural-network output. In our case, at any Q^2, the absolute value of the output should be smaller than C. Hence, for a reliable performance of the network, it is enough to assume that the weights in the linear layer are $|w_{2M+i}| < C \sim a$, $i = 1, \ldots, M + 1$. In analogy to the reasoning in the paragraph above, one can argue that these weights could equally take positive and negative values.

We, therefore, conclude that the prior density for the weights should cover the hypercube $(-a, a)^{3M+1}$ and be symmetric in w_i. This is a rough estimate of the bounds for the model parameters, but the functional form of the prior densities is still arbitrary. In the present analysis, we have considered the Gaussian distribution (15). Such a density profile maximizes the entropy of the system [63]. However, our choice is also supported by further arguments from neural-network computations. In particular, prior (15) modifies the error function (21) through the contribution (16). Such a penalty term has been considered in non-Bayesian approaches [64] to the feedforward neural networks. Bayesian statistics provides a probabilistic justification for (16). Moreover, the Bayesian approach allows to consistently obtain the optimal value of the α. We recall that, in the numerical analysis, the initial value of α is set to $\alpha_0 = 0.001$. Then, the prior Gaussian distribution has a width of $\sqrt{1/\alpha} \approx 30$, which fully covers the region of the parameter space allowed by the constraints.

Eventually, let us remark that the most general Gaussian prior has the form

$$P(\{w_i\} | \mathcal{N}) = \exp \left(- \sum_{i=1}^{3M+1} \frac{\alpha_i}{2} w_i^2 \right), \tag{A2}$$

where every weight w_i has its own regularizer α_i. However, the internal symmetry of the network (P1) allows to reduce the number of regularizers to only four—each for every class of parameters (w_i, w_{M+i}, w_{2M+i}, and w_{3M+1}) with $i = 1 - M$. In Ref. [31], it was verified that the inclusion of more regularizers has a negligible impact on the results but slows down the numerical procedures. Hence, in the present analysis, we consider the simplest and practically very effective scenario.

APPENDIX B: BEST-FIT RESULTS

It is worth noting that each of the sigmoids that constitute the neural networks typically describes a particular feature of the function. If a soft dependence is preferred by the data, some units might be redundant and take very similar values for the weights.

1. BIN0 with deuteron corrections

The best-fit parametrization for the BIN0 data set with the deuteron correction included is

$$\mathcal{N}(Q^2, \{w_j\}) = \frac{w_9}{e^{-(Q^2/Q_0^2)w_1 - w_2} + 1} + \frac{w_{10}}{e^{-(Q^2/Q_0^2)w_3 - w_4} + 1}$$
$$+ \frac{w_{11}}{e^{-(Q^2/Q_0^2)w_5 - w_6} + 1} + \frac{w_{12}}{e^{-(Q^2/Q_0^2)w_7 - w_8} + 1}$$
$$+ w_{13}. \tag{B1}$$

The weights w_{1-13} take the following values:

$$\{w_j\} = \{-2.174\,061, 0.1991\,515, 2.140\,942, -0.1947\,798,$$
$$-2.174\,070, 0.1991\,740, -5.481\,409, 2.501\,837,$$
$$-2.502\,352, 2.308\,397, -2.502\,347,$$
$$3.120\,895, -0.1638\,095\}, \tag{B2}$$

---

[10]At $Q^2 = 0$, we have $f(w_{M+i})$, so it is enough that $|w_{M+i}| < a$.

with a covariance matrix,

$$
A^{-1} = \begin{pmatrix}
3.037317 & -0.1350095 & 0.7164281 & 0.6028375 & -0.9106402 & -0.7254212 & -0.1375870 & 0.1193517 & 1.744848 & 0.7934970 & -1.638952 & -0.6857111 & -0.6863036 \\
-0.1350095 & 6.839114 & 0.5963543 & 2.241853 & -0.7253642 & -2.521010 & 0.08826055 & 0.08952850 & 0.5380971 & -0.4953074 & 0.7737700 & 0.5333276 & 0.3832513 \\
0.7164281 & 0.5963543 & 3.129677 & -0.3601146 & 0.7164453 & 0.5964068 & -0.05068235 & -0.05207627 & 1.066743 & 1.171111 & 1.066819 & 0.6044601 & -1.304295 \\
0.6028375 & 2.241853 & -0.3601146 & 7.284083 & 0.6028388 & 2.242126 & -0.1002227 & -0.05207626 & -0.8354786 & 0.1749902 & -0.8355188 & -0.4839537 & -0.08842895 \\
-0.9106402 & -0.7253642 & 0.7164453 & 0.6028388 & 3.037309 & -0.1350448 & -0.1375865 & 0.1193541 & -1.638966 & 0.7934769 & 1.744873 & -0.6857391 & -0.6862827 \\
-0.7254212 & -2.521010 & 0.5964068 & 2.242126 & -0.1350448 & 6.839396 & 0.08828352 & 0.08953791 & 0.7736800 & -0.4953435 & 0.5381072 & 0.5333633 & 0.3832724 \\
-0.1375870 & 0.08826055 & -0.05068235 & -0.1002227 & -0.1375865 & 0.08828352 & 1.348550 & -0.7524626 & -0.8319859 & 0.4259322 & -0.8320014 & 1.535086 & -0.3655335 \\
0.1193517 & 0.08952850 & -0.05207627 & -0.05207626 & 0.1193541 & 0.08953791 & -0.7524626 & 0.5275705 & 0.6937163 & -0.3336553 & 0.6937285 & -0.9665196 & 0.3068238 \\
1.744848 & 0.5380971 & 1.066743 & -0.8354786 & -1.638966 & 0.7736800 & -0.8319859 & 0.6937163 & 13.95026 & 2.966996 & -6.650710 & -1.081263 & -3.213450 \\
0.7934970 & -0.4953074 & 1.171111 & 0.1749902 & 0.7934769 & -0.4953435 & 0.4259322 & -0.3336553 & 2.966996 & 8.333349 & 2.967110 & 0.5435698 & -8.166890 \\
-1.638952 & 0.7737700 & 1.066819 & -0.8355188 & 1.744873 & 0.5381072 & -0.8320014 & 0.6937285 & -6.650710 & 2.967110 & 13.95042 & -1.081328 & -3.213562 \\
-0.6857111 & 0.5333276 & 0.6044601 & -0.4839537 & -0.6857391 & 0.5333633 & 1.535086 & -0.9665196 & -1.081263 & 0.5435698 & -1.081328 & 2.761672 & -0.4607885 \\
-0.6863036 & 0.3832513 & -1.304295 & -0.08842895 & -0.6862827 & 0.3832724 & -0.3655335 & 0.3068238 & -3.213450 & -8.166890 & -3.213562 & -0.4607885 & 8.103694
\end{pmatrix}. \tag{B3}
$$

As explained in Sec. II B Eq. (8), to obtain $F_A(Q^2)$, function $\mathcal{N}(Q^2, \{w_j\})$ given above should be multiplied by the dipole Eq. (2) with $M_A = 1$ GeV.

2. BIN1 with deuteron corrections

In this case,

$$
\mathcal{N}(Q^2, \{w_j\}) = \frac{w_7}{e^{-(Q^2/Q_0^2)w_1 - w_2} + 1} + \frac{w_8}{e^{-(Q^2/Q_0^2)w_3 - w_4} + 1}
$$
$$
+ \frac{w_9}{e^{-(Q^2/Q_0^2)w_5 - w_6} + 1} + w_{10}, \tag{B4}
$$

with

$$
\{w_j\}_{\mathrm{MP}} = \{-0.0703401, 0.0404197, -0.0703404,
$$
$$
0.0404186, -0.0703372, 0.0404192, 0.299085,
$$
$$
0.299087, 0.299086, 0.554479\}, \tag{B5}
$$

and a covariance matrix,

$$
A^{-1} = \begin{pmatrix}
0.501\,347 & 0.000\,475\,574 & -0.133\,756 & 0.003\,125\,38 & -0.133\,756 & 0.003\,125\,17 & -0.095\,176\,7 & 0.054\,546\,0 & 0.054\,544\,2 & -0.007\,598\,98 \\
0.000\,475\,574 & 0.563\,049 & 0.003\,125\,06 & -0.006\,415\,21 & 0.003\,124\,65 & -0.006\,414\,82 & 0.053\,888\,4 & -0.023\,692\,4 & -0.023\,692\,3 & -0.044\,441\,2 \\
-0.133\,756 & 0.003\,125\,06 & 0.501\,346 & 0.000\,475\,531 & -0.133\,757 & 0.003\,124\,75 & 0.054\,545\,3 & -0.095\,176\,5 & 0.054\,544\,8 & -0.007\,598\,98 \\
0.003\,125\,38 & -0.006\,415\,21 & 0.000\,475\,531 & 0.563\,050 & 0.003\,124\,57 & -0.006\,414\,96 & -0.023\,692\,8 & 0.053\,888\,2 & -0.023\,691\,8 & -0.044\,441\,5 \\
-0.133\,756 & 0.003\,124\,65 & -0.133\,757 & 0.003\,124\,57 & 0.501\,347 & 0.000\,474\,719 & 0.054\,545\,8 & 0.054\,545\,7 & -0.095\,178\,3 & -0.007\,598\,68 \\
0.003\,125\,17 & -0.006\,414\,82 & 0.003\,124\,75 & -0.006\,414\,96 & 0.000\,474\,719 & 0.563\,050 & -0.023\,692\,8 & -0.023\,692\,1 & 0.053\,888\,8 & -0.044\,441\,5 \\
-0.095\,176\,7 & 0.053\,888\,4 & 0.054\,545\,3 & -0.023\,692\,8 & 0.054\,545\,8 & -0.023\,692\,8 & 0.505\,143 & -0.101\,488 & -0.101\,488 & -0.154\,622 \\
0.054\,546\,0 & -0.023\,692\,4 & -0.095\,176\,5 & 0.053\,888\,2 & 0.054\,545\,7 & -0.023\,692\,1 & -0.101\,488 & 0.505\,144 & -0.101\,488 & -0.154\,623 \\
0.054\,544\,2 & -0.023\,692\,3 & 0.054\,544\,8 & -0.023\,691\,8 & -0.095\,178\,3 & 0.053\,888\,8 & -0.101\,488 & -0.101\,488 & 0.505\,144 & -0.154\,623 \\
-0.007\,598\,98 & -0.044\,441\,2 & -0.007\,598\,98 & -0.044\,441\,5 & -0.007\,598\,68 & -0.044\,441\,5 & -0.154\,622 & -0.154\,623 & -0.154\,623 & 0.246\,587
\end{pmatrix}. \tag{B6}
$$

[1] L. Alvarez-Ruso *et al.*, Prog. Part. Nucl. Phys. **100**, 1 (2017).

[2] C. Patrignani *et al.* (Particle Data Group), Chin. Phys. C **40**, 100001 (2016).

[3] M. González-Alonso, O. Naviliat-Cuncic, and N. Severijns, Prog. Part. Nucl. Phys. **104**, 165 (2018).

[4] J. Arrington, C. D. Roberts, and J. M. Zanotti, J. Phys. G: Nucl. Part. Phys. **34**, S23 (2007).

[5] T. Katori and M. Martini, J. Phys. G: Nucl. Part. Phys. **45**, 013001 (2018).

[6] W. A. Mann *et al.*, Phys. Rev. Lett. **31**, 844 (1973).

[7] S. J. Barish, J. Campbell, G. Charlton, Y. Cho, M. Derrick, R. Engelmann, L. G. Hyman, D. Jankowski, A. Mann, B. Musgrave, P. Schreiner, P. F. Schultz, R. Singer, M. Szczekowski, T. Wangler, H. Yuta, V. E. Barnes, D. D. Carmony, A. F. Garfinkel, and G. M. Radecky, Phys. Rev. D **16**, 3103 (1977).

[8] K. L. Miller, S. J. Barish, A. Engler, R. W. Kraemer, B. J. Stacey, M. Derrick, E. Fernandez, L. Hyman, G. Levman, D. Koetke, B. Musgrave, P. Schreiner, R. Singer, A. Snyder, S. Toaff, D. D. Carmony, G. M. Radecky, V. E. Barnes, A. F. Garfinkel, R. Ammar, D. Coppage, D. Day, R. Davis, N. Kwak, and R. Stump, Phys. Rev. D **26**, 537 (1982).

[9] N. J. Baker, A. M. Cnops, P. L. Connolly, S. A. Kahn, H. G. Kirk, M. J. Murtagh, R. B. Palmer, N. P. Samios, and M. Tanaka, Phys. Rev. D **23**, 2499 (1981).

[10] T. Kitagaki, H. Yuta, S. Tanaka, A. Yamaguchi, K. Abe, K. Hasegawa, K. Tamai, H. Sagawa, K. Akatsuka, K. Furuno, K. Tamae, M. Higuchi, M. Sato, S. A. Kahn, M. J. Murtagh, R. B. Palmer, N. P. Samios, and M. Tanaka, Phys. Rev. D **42**, 1331 (1990).

[11] T. Kitagaki, S. Tanaka, H. Yuta, K. Abe, K. Hasegawa, A. Yamaguchi, K. Tamai, T. Hayashino, Y. Otani, H. Hayano, H. Sagawa, R. A. Burnstein, J. Hanlon, H. A. Rubin, C. Y. Chang, S. Kunori, G. A. Snow, D. Son, P. H. Steinberg, D. Zieminska, R. Engelmann, T. Kafka, S. Sommars, C. C. Chang, W. A. Mann, A. Napier, and J. Schneps, Phys. Rev. D **28**, 436 (1983).

[12] D. Allasia *et al.*, Nucl. Phys. B **343**, 285 (1990).

[13] A. Bodek, S. Avvakumov, R. Bradford, and H. S. Budd, Eur. Phys. J. C **53**, 349 (2008).

[14] B. Bhattacharya, R. J. Hill, and G. Paz, Phys. Rev. D **84**, 073006 (2011).

[15] B. Bhattacharya, G. Paz, and A. J. Tropiano, Phys. Rev. D **92**, 113011 (2015).

[16] A. S. Meyer, M. Betancourt, R. Gran, and R. J. Hill, Phys. Rev. D **93**, 113015 (2016).

[17] R. J. Hill, P. Kammel, W. J. Marciano, and A. Sirlin, Rept. Prog. Phys. **81**, 096301 (2018).

[18] E. Berkowitz *et al.*, arXiv:1704.01114.

[19] C. Alexandrou, M. Constantinou, K. Hadjiyiannakou, K. Jansen, C. Kallidonis, G. Koutsou, and A. Vaquero Aviles-Casco, Phys. Rev. D **96**, 054507 (2017).

[20] S. Capitani, M. Della Morte, D. Djukanovic, G. M. von Hippel, J. Hua, B. Jäger, P. M. Junnarkar, H. B. Meyer, T. D. Rae, and H. Wittig, Int. J. Mod. Phys. A **34**, 1950009 (2019).

[21] R. Gupta, Y.-C. Jang, H.-W. Lin, B. Yoon, and T. Bhattacharya, Phys. Rev. D **96**, 114503 (2017).

[22] D.-L. Yao, L. Alvarez-Ruso, and M. J. Vicente-Vacas, Phys. Rev. D **96**, 116022 (2017).

[23] K. M. Graczyk and C. Juszczak, Phys. Rev. C **90**, 054334 (2014).

[24] R. D. Ball, S. Forte, A. Guffanti, E. R. Nocera, G. Ridolfi, and J. Rojo (NNPDF Collaboration), Nucl. Phys. B **874**, 36 (2013).

[25] C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, 1995).

[26] K. Hornik, M. Sinchcombe, and W. Halbert, Neural Networks **2**, 359 (1989).

[27] G. D'Agostini, *Bayesian Reasoning in Data Analysis* (World Scientific, Singapore, 2003).

[28] L. De Cruz, T. Vrancx, P. Vancraeyveld, and J. Ryckebusch, Phys. Rev. Lett. **108**, 182002 (2012).

[29] J. D. McDonnell, N. Schunck, D. Higdon, J. Sarich, S. M. Wild, and W. Nazarewicz, Phys. Rev. Lett. **114**, 122501 (2015).

[30] D. MacKay, Bayesian methods for adaptive models, Ph.D. thesis, California Institute of Technology, 1991.

[31] K. M. Graczyk, P. Plonski, and R. Sulej, J. High Energy Phys. 09 (2010) 053.

[32] K. M. Graczyk, Phys. Rev. C **84**, 034314 (2011).

[33] K. M. Graczyk, Phys. Rev. C **88**, 065205 (2013).

[34] K. M. Graczyk and C. Juszczak, J. Phys. G: Nucl. Part. Phys. **42**, 034019 (2015).

[35] K. M. Graczyk and C. Juszczak, Phys. Rev. C **91**, 045205 (2015).

[36] F. Rosenblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).

[37] G. Cybenko, Math Control, Signal **2**, 303 (1989).

[38] K.-I. Funahashi, Neural Networks **2**, 183 (1989).

[39] R. Hecht-Nielsen, *International 1989 Joint Conference on Neural Networks*, Washington, DC, 1989 (IEEE, New York, 1989), Vol. 1, pp. 593–605.

[40] N. E. Cotter, IEEE Trans, Neural Networks **1**, 290 (1990).

[41] Y. Ito, Neural Networks **4**, 385 (1991).

[42] V. Y. Kreinovich, Neural Networks **4**, 381 (1991).

[43] S. Geman, E. Bienenstock, and R. Doursat, Neural Comput. **4**, 1 (1992).

[44] J. O. Berger and W. H. Jefferys, J. Italian Statist. Soc. **1**, 17 (1992).

[45] W. H. Jefferys and J. O. Berger, Am. Sci. **80**, 64 (1992).

[46] R. Pohl, A. Antognini, F. Nez, F. D. Amaro, F. Biraben *et al.*, Nature (London) **466**, 213 (2010).

[47] I. Sick, Atoms **6**, 2 (2018).

[48] D. J. C. MacKay, Neural Comput. **4**, 415 (1992).

[49] D. J. C. MacKay, Neural Comput. **4**, 448 (1992).

[50] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, New York, 1985).

[51] C. H. Llewellyn Smith, Phys. Rep. **3**, 261 (1972).

[52] S. Galster, H. Klein, J. Moritz, K. H. Schmidt, D. Wegener, and J. Bleckwenn, Nucl. Phys. B **32**, 221 (1971).

[53] J. Nieves, J. E. Amaro, and M. Valverde, Phys. Rev. C **70**, 055503 (2004); **72**, 019902(E) (2005).

[54] S. K. Singh and H. Arenhövel, Z. Phys. A **324**, 347 (1986).

[55] S. J. Barish, M. Derrick, T. Dombeck, L. G. Hyman, K. Jaeger, B. Musgrave, P. Schreiner, R. Singer, A. Snyder, V. E. Barnes, D. D. Carmony, and A. F. Garfinkel, Phys. Rev. D **19**, 2521 (1979).

[56] G. D'Agostini, Nucl. Instrum. Methods Phys. Res., Sect. A **346**, 306 (1994).

[57] K. M. Graczyk, D. Kielczewska, P. Przewlocki, and J. T. Sobczyk, Phys. Rev. D **80**, 093001 (2009).

[58] K. Levenberg, Quart. Appl. Math. **2**, 164 (1944).

[59] D. W. Marquardt, J. Soc. Indust. Appl. Math. **11**, 431 (1963).

[60] G. Shen, L. E. Marcucci, J. Carlson, S. Gandolfi, and R. Schiavilla, Phys. Rev. C **86**, 035503 (2012).

[61] O. Moreno, T. W. Donnelly, J. W. Van Orden, and W. P. Ford, Phys. Rev. D **92**, 053006 (2015).

[62] http://www.wcss.wroc.pl.

[63] G. P. Lepage, B. Clark, C. T. H. Davies, K. Hornbostel, P. B. Mackenzie, C. Morningstar, and H. Trottier, Nucl. Phys. Proc. Suppl. **106**, 12 (2002).

[64] G. E. Hinton, in *PARLE Parallel Architectures and Languages Europe*, edited by J. W. de Bakker, A. J. Nijman, and P. C. Treleaven (Springer, Berlin/Heidelberg, 1987), pp. 1–13.