# Assessing Correlated Truncation Errors in Modern Nucleon-Nucleon Potentials

P. J. Millican,* R. J. Furnstahl,† and J. A. Melendez‡
*Department of Physics, The Ohio State University, Columbus, OH 43210, USA*

D. R. Phillips§
*Department of Physics & Astronomy and Institute of Nuclear & Particle Physics, Ohio University, Athens, OH 45701, USA and Department of Physics, Chalmers University of Technology, SE-41296 Göteborg, Sweden*

M. T. Pratola¶
*Department of Statistics, The Ohio State University, Columbus, OH 43210, USA*

(Dated: February 21, 2024)

We test the BUQEYE model of correlated effective field theory (EFT) truncation errors on Reinert, Krebs, and Epelbaum's semi-local momentum-space implementation of the chiral EFT ($\chi$EFT) expansion of the nucleon-nucleon (NN) potential. This Bayesian model hypothesizes that dimensionless coefficient functions extracted from the order-by-order corrections to NN observables can be treated as draws from a Gaussian process (GP). We combine a variety of graphical and statistical diagnostics to assess when predicted observables have a $\chi$EFT convergence pattern consistent with the hypothesized GP statistical model. Our conclusions are: First, the BUQEYE model is generally applicable to the potential investigated here, which enables statistically principled estimates of the impact of higher EFT orders on observables. Second, parameters defining the extracted coefficients such as the expansion parameter $Q$ must be well chosen for the coefficients to exhibit a regular convergence pattern—a property we exploit to obtain posterior distributions for such quantities. Third, the assumption of GP stationarity across lab energy and scattering angle is not generally met; this necessitates adjustments in future work. We provide a workflow and interpretive guide for our analysis framework, and show what can be inferred about probability distributions for $Q$, the EFT breakdown scale $\Lambda_b$, the scale associated with soft physics in the $\chi$EFT potential $m_{\rm eff}$, and the GP hyperparameters. All our results can be reproduced using a publicly available Jupyter notebook, which can be straightforwardly modified to analyze other $\chi$EFT NN potentials.



## I. INTRODUCTION

Nucleon-nucleon (NN) potentials based on a chiral effective field theory (EFT) expansion, suitably augmented by three-nucleon forces, have found fruitful application in computations of finite nuclei and of nuclear and neutron matter [1–5]. However, given that these nuclear forces are derived from an EFT at finite order, we need to know about the error induced by the truncation of that expansion. This is essential for robust uncertainty quantification (UQ) because such errors are expected to be at least comparable to uncertainties induced by fitting the rather accurate data in the NN sector and precisely known binding energies of particular nuclides. Failure to account for these truncation errors—including their correlations in energy and angle—could lead to bias and prevent us from properly propagating uncertainties to predictions. In this paper, we test the BUQEYE model of correlated EFT truncation errors.

Various chiral NN potentials have been developed recently [6–13]. These potentials differ in the way they are regulated (local, non-local, or a mix of the two called semi-local), the value of the associated regulator parameter, the order to which the EFT expansion is carried out (labeled $\mathrm{N^2LO}$, $\mathrm{N^3LO}$, $\mathrm{N^4LO}$, and even beyond; see below), and the data to which the parameters in the potentials are fit. In some cases the potentials include explicit $\Delta(1232)$ degrees of freedom. However, all claim to be an implementation of the idea—originally proposed by Weinberg over thirty years ago—that the NN potential can be organized as an expansion in powers of the parameter $Q \equiv f(p_{\rm rel}, m_\pi)/\Lambda_b$, where $p_{\rm rel}$ is the relative momentum of the two nucleons, $m_\pi$ is the rest mass of the pion, and $\Lambda_b$ is the breakdown scale of the theory) [14, 15]. The benefit of such an organization is that systematically smaller effects occur at higher order in the expansion, and so calculations of observables should become progressively more accurate when potentials of higher order are employed.

In Refs. [16, 17] we proposed a pointwise Bayesian statistical model (building on Refs. [6, 18]) to estimate EFT truncation errors for predicted observables $y$. "Pointwise" in this context means that the error model is applied independently at each energy or angle in the input space (generically denoted $x$). The model incorporates expert knowledge [19] on the convergence pattern (inherited from the EFT power counting) into prior distributions, and subsequently updates these beliefs given order-by-order predictions of observables $\{y_n\}$, thereby formalizing the notion of EFT convergence for $y$.

* millican.7@osu.edu
† furnstahl.1@osu.edu
‡ melendez.27@osu.edu
§ phillid1@ohio.edu
¶ mpratola@stat.osu.edu

Subsequently, we extended this pointwise model to a curvewise treatment of dimensionless coefficient *functions* $c_n(x)$ extracted from the order-by-order predictions $\{y_n(x)\}$ [20]. The BUQEYE model postulates that the $c_n(x)$ can be modeled as random draws from a Gaussian process (GP). GPs are powerful tools for both regression and classification, and have become popular in statistics, physics, applied mathematics, machine learning, and geostatistics [21–23]. The BUQEYE model uses them in an atypical way by learning the GP parameters characterizing the distribution of the known coefficient functions and, by induction, using this same GP to predict error bands for unknown higher-order contributions. This correlated truncation error model has been applied to infinite nuclear matter [24, 25], nucleon-nucleus elastic scattering [26], $np \leftrightarrow d\gamma$ reactions [27], muon capture by deuterium [28], experimental design for proton Compton scattering [29], and parameter estimation of NN low-energy constants (LECs) [30] (see also [31–34]).

In this paper, we exemplify a workflow for using the BUQEYE correlated error model. We apply it to examine in detail whether the model can accurately describe order-by-order convergence of the observables predicted by modern NN potentials. However, in order to achieve this the input space and hyperparameters of the GP that encodes our error model must be carefully chosen. We limit ourselves to the neutron-proton ($np$) scattering observables predicted using the semi-local momentum-space (SMS) potential of Reinert, Krebs, and Epelbaum with cutoff 500 MeV [11]. (Other potentials and proton-proton scattering will be examined in future work.) We emphasize that we do not refit the LECs of this potential to data here; instead, the LECs for the N$^k$LO potential are taken from the N$^k$LO fit as given by the original publication [11]. However, because our truncation-error model can be embedded within a Bayesian parameter estimation framework (see Ref. [31]), we are setting the stage for a full Bayesian treatment of such potentials with correlated errors.

In Sec. II we review the GP model of EFT truncation errors [20], starting with the extraction of the coefficient functions from order-by-order predictions of the total cross section based on a chosen $\chi$EFT potential. With appropriate choices of input parameters, visual inspection of these functions suggests they are consistent with statistical draws from a common GP. The tools for a more complete analysis that uses complementary statistical diagnostics and outputs of our GP analysis are summarized in Sec. III, which includes a schematic workflow for model-checking. We then extract posteriors for the expansion parameter $Q$ in Sec. IV and thereby expose the inadequacy of a GP model that is stationary in an input space defined by the scattering angle and energy. This flaw is mitigated by the requirement of statistical consistency for the EFT's error bands, which implies Bayesian credibility intervals for both the EFT breakdown momentum $\Lambda_b$ and the effective soft scale $m_{\text{eff}}$ (typically identified with the pion mass for $\chi$EFT). We show that consistent choices for these EFT scales produce approximate stationarity provided we employ a suitable input space and consider only observables corresponding to relative momenta above the pion mass. In Sec. V we demonstrate how this informed choice of input parameters leads to coefficient functions consistent with the BUQEYE model across observables. Two representative applications are given in Sec. VI before Sec. VII provides a summary and highlights outstanding problems and future work.

## II. A MODEL OF EFT CORRELATED TRUNCATION ERRORS

We recapitulate briefly in Sec. II A the details of the BUQEYE model for correlated truncation errors [20]. For our analysis we need to convert a potential's order-by-order predictions to dimensionless coefficients, which requires choosing a set of parametrizations. These are introduced and their rationales and functional forms explained in Sec. II B. Then, in Sec. II C we give a simple example of the impact that these parametrization choices can have and discuss how the convergence pattern explained in Sec. II A can be enhanced for a given set of parametrization choices.

### A. Recap of BUQEYE model

We follow Refs. [16, 20] and formalize the power counting of the EFT by writing an observable $y$ as

$$y_k(x) = y_{\text{ref}}(x) \sum_{n=0}^{k} c_n(x) Q^n(x), \tag{1}$$

where $k$ is the highest computed order of the EFT, $Q$ is the dimensionless expansion parameter, and $x$ is the input space(s). In the case of angular NN scattering observables $x = (E_{\text{lab}}, \theta)$, the lab-frame energy and center-of-mass scattering angle, respectively, whereas $x = E_{\text{lab}}$ for the total cross section $\sigma_{\text{tot}}$. The dimensionful reference scale $y_{\text{ref}}$ is taken to be $y_{\text{ref}}(x) \equiv y_k(x)$ for the total and differential cross sections and $y_{\text{ref}}(x) \equiv 1$ for spin observables, except in Sec. V D where an example of a mischosen $y_{\text{ref}}$ is provided. The theoretical uncertainty, $\delta y_k$, is then written as:

$$\delta y_k(x) = y_{\text{ref}}(x) \sum_{n=k+1}^{\infty} c_n(x) Q^n(x). \tag{2}$$

Given choices of $y_{\text{ref}}$, the expansion parameter $Q$ and the input space $x$, the observable coefficients $c_n$ for $n = 0, \ldots, k$ are completely determined by the order-by-order predictions $y_n$.

The formulation of Eqs. (1) and (2) allows us to estimate the size and correlations of the truncation error $\delta y_k$ if we can relate the properties of these low-order $c_n$s

to the properties of the higher-order $c_n$s. An underlying assumption of the BUQEYE model is that the $c_n$, if properly extracted, should share common properties due to the regularity with which a well-constructed EFT should converge. For most NN observables the $\chi$EFT expansion does not converge monotonically from above or below. The coefficients thus show no preference for positive or negative values and their mean is about zero. The smoothness of the EFT corrections as functions of $x$ is also inherited by the coefficients. But the particular choices made for $Q$, $x$, and $y_{\rm ref}$ can affect whether the $c_n$ have similar properties across the domain; this is the property of *stationarity*.

If we can show that the known $c_n$ are effectively random and embody similar properties, we can learn these properties to inform ourselves of what higher-order $c_n$ should look like. To this end, we hypothesize that the $c_n$ are independent and identically distributed (iid) draws from a Gaussian process[1]

$$c_n(x)\,|\,\boldsymbol{\theta} \overset{\text{iid}}{\sim} \mathcal{GP}[\mu=0, \bar{c}^2 r(x,x';L)], \tag{3}$$

which is a collection of random variables, any finite number of which have a joint Gaussian distribution [35]. We take the mean function $\mu$ to be identically zero. The *a priori* unknown GP parameters are $\boldsymbol{\theta} = \{\bar{c}^2, L\}$,[2] where $\bar{c}$, the marginal standard deviation, controls the average magnitude of the curves, and $L$, the correlation length scale matrix, controls the approximate frequencies with which the curves oscillate. The correlation function $r$ governs the smoothness properties of the curves.

Our approach to modeling truncation errors is then based on the idea that from the knowledge of a few coefficients, we can create a statistically meaningful distribution for the truncation error $\delta y_k(x)$. In particular, Ref. [20] showed that, given (3),

$$\delta y_k(x)\,|\,\boldsymbol{\theta}, Q \sim \mathcal{GP}[0, \bar{c}^2 R_{\delta k}(x,x';L)], \tag{4}$$

where

$$R_{\delta k}(x,x';L) \equiv y_{\rm ref}(x) y_{\rm ref}(x') \frac{[Q(x)Q(x')]^{k+1}}{1-Q(x)Q(x')} r(x,x';L). \tag{5}$$

Given a prior $\text{pr}(\boldsymbol{\theta})$, the $\boldsymbol{\theta}$ needed to define Eq. (4) can be estimated—and subsequently integrated over, if desired—by using the low-order $c_n$ (see Ref. [20]).

To match the observed smoothness of calculated coefficient functions, we choose the squared exponential (a.k.a. radial basis function or Gaussian) for $r$ (see Table I). But this choice also assumes stationarity, which will prove to be questionable for the energy/momentum dependence of many of the coefficients. The length scale prior is taken to be uninformative: a uniform distribution over all positive values (an improper prior, but one that nonetheless expresses the wide range of values that the length scale can take). In addition, we use a weakly informative conjugate prior for the $c_n$ variance that accurately reflects our prior beliefs: If the $c_n$ are naturally sized, we expect that $\bar{c} \approx 1$, but allow it to vary if the data support other values. We have tested prior dependence in previous works [16, 17] and found it to be slight.[3]

Our statistical model can also incorporate symmetry constraints on the values of scattering observables [20]. In particular, some spin observables are constrained to take the value of zero at particular angles when time-reversal invariance is imposed as a necessary symmetry. Specifically, time-reversal symmetry fixes $A(E_{\rm lab}, \theta = 0^\circ) = 0$ and $A_y(E_{\rm lab}, \theta = \{0^\circ, 180^\circ\}) = 0$ for all $E_{\rm lab}$ [36], where we follow the nomenclature of Ref. [17]. These symmetry constraints shrink the uncertainty band in the immediate vicinity of the constraint point(s); for how this can impact statistical diagnostics, see Sec. V C.

### B. Options for parametrization

Inverting Eq. (1) enables the extraction from NN observable data of coefficient functions given some choice of $y_{\rm ref}$ and $Q$ parametrization, and these functions can be plotted against an input space $x$. Then, we assess those coefficients visually and statistically to see if our choices of $Q$, $y_{\rm ref}$, and $x$ yield coefficients that are stationary across the domain of interest and manifest expected power counting by exhibiting common properties across orders. Here we focus on choices for $Q$ and $x$.

The conventional form of $Q(p, m_{\rm eff})$ [6] is

$$Q_{\rm max}(p, m_{\rm eff}) = \frac{\max(p, m_{\rm eff})}{\Lambda_b}\ , \tag{6}$$

with $p = p_{\rm rel}$ and $m_{\rm eff} = m_\pi = 138\,\text{MeV}$. This choice of $Q$ prescription was made in previous work (e.g., in [17, 31]). $Q_{\rm max}$ can pose issues for GP fitting due to the cusp at $p_{\rm rel} = m_{\rm eff}$, so we introduce a differentiable smooth-maximum ("smoothmax" or "smax") version given by [17]

$$Q_{\rm smax}(p, m_{\rm eff}) = \frac{1}{\Lambda_b} \frac{p^i + m_{\rm eff}^i}{p^{i-1} + m_{\rm eff}^{i-1}}\ , \tag{7}$$

where we choose $i = 8$ (see Fig. 2 for a comparison).

---

[1] The $z \sim \cdots$ notation is a common shorthand in statistical literature for "$z$ is distributed as." We use "$\text{pr}(z) = \cdots$" as well. Also, "$z\,|\,I$" is read as "$z$ given $I$."

[2] We follow statistical convention and use $\boldsymbol{\theta}$ to denote the vector of these parameters but also use $\theta$ for the center-of-mass scattering angle.

[3] The insensitivity to the precise form of the prior motivates the use of conjugate priors, for which updating from prior to posterior as we accumulate data on low-order coefficients is analytic. The posterior for $\mu, \bar{c}^2$ has the same functional form as the prior (see Ref. [20] for additional discussion). There are no conjugate priors for $L$ or $Q$.

In Figs. 1(a) and 1(b), we compare coefficient functions for the $np$ total cross section ($\sigma_{\rm tot}$) for $Q_{\rm smax}$ and two choices of the input space, $x = E_{\rm lab}$ and $x = p_{\rm rel}$. Here,

$$E_{\rm lab} = \frac{p_{\rm rel}^2 - m_1 m_2 + \sqrt{(p_{\rm rel}^2 + m_1^2)(p_{\rm rel}^2 + m_2^2)}}{m_2}, \quad (8)$$

where $m_1$ is the mass of the beam particle and $m_2$ the mass of the target particle in MeV, which simplifies to

$$E_{\rm lab} = \frac{2p_{\rm rel}^2}{M} \quad (9)$$

in the case where $m_1 = m_2 = M$. This choice of independent variable $x$ can cause the $c_n$ to have a shorter local wavelength at low $E_{\rm lab}$ compared to high $E_{\rm lab}$.

As seen in Figs. 1(a) and 1(b), choosing to parametrize $Q$ with $Q_{\rm smax}$ (or $Q_{\rm max}$) can lead to the $c_n$ growing systematically with $n$ for $p_{\rm rel} \lesssim m_\pi$. This behavior would violate the BUQEYE model's assumption that the coefficients share common properties across orders. A third prescription,

$$Q_{\rm sum}(p, m_{\rm eff}) = \frac{p + m_{\rm eff}}{\Lambda_b + m_{\rm eff}}, \quad (10)$$

was originally devised to ameliorate this issue as seen in Figs. 1(c) and 1(d); a more direct motivation for Eq. (10) will be given in Sec. IV.

Both $Q_{\rm smax}$ and $Q_{\rm sum}$ are defined so that the breakdown scale $\Lambda_b$ is defined by $Q(p = \Lambda_b, m_{\rm eff}) = 1$. However, their differing functional forms should be borne in mind when comparing values of $\Lambda_b$ and $m_{\rm eff}$ between the two prescriptions; more details are given in Sec. IV. We note that none of these three forms for $Q$ are based on analytical arguments regarding the combinations of $p$ and $m_\pi$ that appear in $\chi$EFT Feynman diagrams.

Additionally, we must choose a functional form for the characteristic momentum $p$. We consider three options for the momentum scale that appears in the expansion parameter: $p = p_{\rm rel}$; $p = q_{\rm CM}$, where

$$q_{\rm CM}^2 = (\vec{p} - \vec{p}\,')^2 \;\Rightarrow\; q_{\rm CM} = p_{\rm rel}\sqrt{2(1 - \cos\theta)} \quad (11)$$

when $p = p' = p_{\rm rel}$; and a combination of the two, $p = p_{\rm smax}(p_{\rm rel}, q_{\rm CM})$, where

$$p_{\rm smax}(x, y) = \frac{1}{N}\log_{1.01}(1.01^{Nx} + 1.01^{Ny}) \quad (12)$$

with $N = 5$, which is a smooth maximum interpolation function borrowed from deep-learning applications [37].

Besides parametrizations of $Q$ and $p$, we have also tested parametrizations of the input space $x = (x_E, x_\theta)$, where $x_E$ is the energy-dependent input space and $x_\theta$ the angle-dependent one. Which options for $x$ are available depends upon which physical quantities the observables depend upon. The observables considered in this paper are the total cross section $\sigma_{\rm tot}$; the differential cross section; and the spin observables $D$, $A_{xx}$, $A_{yy}$, $A$, and $A_y$, the latter of which is elsewhere referred to as $P$ or

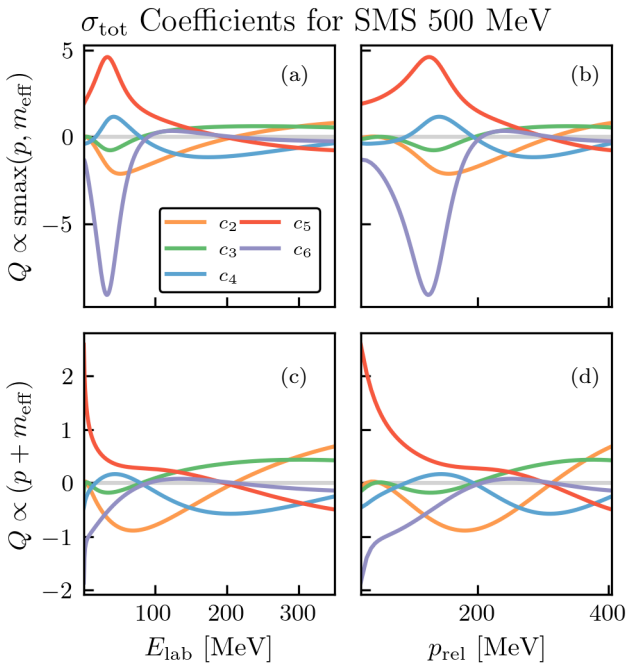



FIG. 1. Observable coefficients for the total neutron-proton cross section ($\sigma_{\rm tot}$), under various assumptions for $Q(p)$ and the input space $x$. Predictions are generated with the 500 MeV SMS potential from Ref. [11] and assume a fixed value of $m_{\rm eff} = 138$ MeV (i.e., the approximate pion rest mass) and $\Lambda_b = 600$ MeV. The top row uses a smoothed maximum $Q_{\rm smax}$ [Eq. (7))] while the bottom row uses $Q_{\rm sum}$ [Eq. (10)]. The left column uses $E_{\rm lab}$ as the $x$-variable while the right column uses $p_{\rm rel}$. (The relationship between the two is given by Eq. (9).) Choosing $x = p_{\rm rel}$ and $Q = Q_{\rm sum}$ results in coefficients that look the most stationary, i.e., similarly sized and with similar length scale across the domain and across coefficients.

$PB$ [10, 17, 36, 38]. All are functions of scattering angle and lab energy except for the total cross section, which is a function of lab energy alone. When the observable depends on both quantities, it is of course also possible to plot that observable (or the dimensionless coefficients derived therefrom) at some fixed lab energy or scattering angle against the remaining other quantity, which is allowed to vary. In the fixed-angle case, the two options for $x_E$ are those discussed already, namely $E_{\rm lab}$ and $p_{\rm rel}$. In the fixed-energy case, there are four options for parametrizing $x_\theta$: $\theta$, $-\cos(\theta)$, $q_{\rm CM}$, and $q_{\rm CM}^2$.

After exploring many parametrizations for use with the assumed stationary GPs (case studies for $Q(p)$, and $(x_E, x_\theta)$ can be found in Secs. V A and V B), we have collected our preferred choices for NN observable-specific and GP quantities in Table I. These are the same physically grounded choices we use to generate many of the figures in Secs. IV and V, where we show that—with additional restriction on the input space—they lead in many (but not all) cases to coefficient functions manifesting the model assumptions adopted here: naturalness and stationarity. A more complete implementation of the cor-

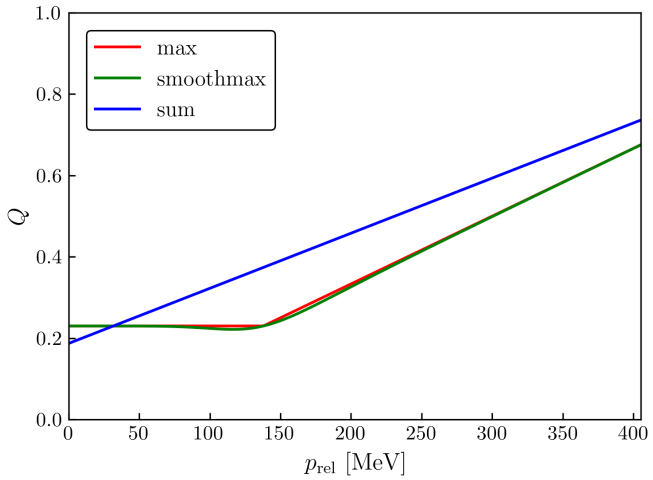



FIG. 2. Plot of the three parametrizations of the dimensionless expansion parameter $Q$ that are tested in this paper — $Q_{\rm max}$, $Q_{\rm smax}$, and $Q_{\rm sum}$ — versus relative momentum $p_{\rm rel}$. In plotting all three prescriptions here, it is assumed that $\Lambda_b = 600\,\text{MeV}$ and $m_{\rm eff} = m_\pi = 138\,\text{MeV}$, but that is not the case elsewhere in this work.

| NN Observable-Specific Quantities | | | |
|---|---|---|---|
| Quantity | $\sigma_{\rm tot}$ | $\sigma(\theta)$ | Spin |
| $y_{\rm ref}$ | $y_k$ | $y_k$ | 1 |
| $(x_E,\, x_\theta)$ | $(p_{\rm rel})$ | $(p_{\rm rel}, -\cos(\theta))$ | $(p_{\rm rel}, -\cos(\theta))$ |
| $L$ | $\ell_E$ | $\text{diag}(\ell_E, \ell_\theta)$ | $\text{diag}(\ell_E, \ell_\theta)$ |
| **Generic Gaussian Process Quantities for NN** | | | |
| Characteristic Momentum | $p = p_{\rm rel}$ | | |
| Expansion Parameter | $Q = (p+m)/(\Lambda_b + m)$ | | |
| Correlation Function | $r(x,x';L) = e^{-\frac{1}{2}(x-x')^\intercal L^{-1}(x-x')}$ | | |
| Variance Prior | $\bar{c}^2 \sim \chi^{-2}(\nu_0 = 1, \tau_0^2 = 1^2)$ | | |
| Length Scale Prior | $\text{pr}(\ell_i) \propto 1,\ \ell_i > 0$ | | |

TABLE I. Preferred assumptions about the model parametrization choices for each NN observable, under the constraint of stationary GPs. (See Sec. III of Ref. [20] for further details.) The correlation function is the squared exponential, which assumes that the $c_n$ are very smooth and stationary. The GP variance prior is a weakly informative inverse chi-squared distribution, while the length scales priors are (positive) uninformative and uniform. The angular observables live in a two-dimensional (2D) space, so the matrix of correlation lengths, $L$, is a 2D diagonal matrix with $\ell_E$ and $\ell_\theta$ on the diagonal.

related error model will require future generalizations to nonstationary GPs that can provide the additional flexibility needed to accommodate the structure observed in NN data.

### C. Varying the breakdown scale and effective mass

As we have discussed, Fig. 1 shows $c_n(x)$ extracted with two choices of $Q$ (namely $Q_{\rm smax}$ and $Q_{\rm sum}$), plotted against two choices of independent variables $x$ (namely $E_{\rm lab}$ and $p_{\rm rel}$) for the $np$ $\sigma_{\rm tot}$ calculated with the 500 MeV cutoff potential from Ref. [11]. In order to comport with our model, the coefficients should exhibit *naturalness*, the tendency to have variances not too much different than unity across orders. Additionally, they should exhibit *stationarity*, the tendency to have roughly the same length scale and variance (or magnitude, in this case) across the input space. When these qualities of the coefficients are not in evidence, an informed change in the parametrization of $Q(p)$, $y_{\rm ref}$, and/or $x$ can sometimes turn a failure of our model into a success.

Here, the behavior of the coefficients in Figs. 1(a) and 1(b) at low energies/momenta suggests that the coefficients behave unusually there. Specifically, compared to elsewhere in the input space, the length scale appears shorter and the variance larger. But with the choice of $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = m_\pi, \Lambda_b = 600\,\text{MeV})$ and $x = p_{\rm rel}$ (see Fig. 1(d)), then all $c_n$ appear more regular, except possibly at very small $p_{\rm rel}$ and for rather high-order coefficients. (The putative failure there is perhaps because specifics of the fitting procedure used in Ref. [11] get amplified by the smallness of $O(Q^5)$ and $O(Q^6)$ contributions, or because the operative power-counting scheme for that low-momentum regime is that of pionless EFT.) Even in Fig. 1(d), however, we notice discrepancies between the exhibited and ideal order-by-order convergence in the coefficients. The question is, then: Can we do better?

In Ref. [39], an effective value of 200–225 MeV for the soft scale $m_{\rm eff}$, which differs from the heretofore assumed value of 138 MeV, was extracted from credible-interval ("weather") plots (see Sec. III C) based upon $\sigma_{\rm tot}$. When we make this change, we observe (see Fig. 3) in the coefficients much stronger evidence of the features we hope to see — namely, naturalness and stationarity in the length scale. Nevertheless, the region of $p_{\rm rel} < m_\pi$ remains problematic.

The visible improvements between Fig. 1 and Fig. 3 suggest the possibility that the values of $m_{\rm eff}$ and $\Lambda_b$ can be chosen to optimize the convergence pattern of a potential under a given parametrization (such as the choices of $Q(p)$ and $x$). By means of a Bayesian-statistical approach to parameter estimation (outlined in Sec. III E), we can extract a maximum *a posteriori* (MAP) value for these parameters and re-plot the coefficients *mutatis mutandis*. A further discussion of this possibility, with accompanying figures, is given in Sec. IV.

One final note: There may be a region in the input space that is resistant to our efforts to obtain stationar-

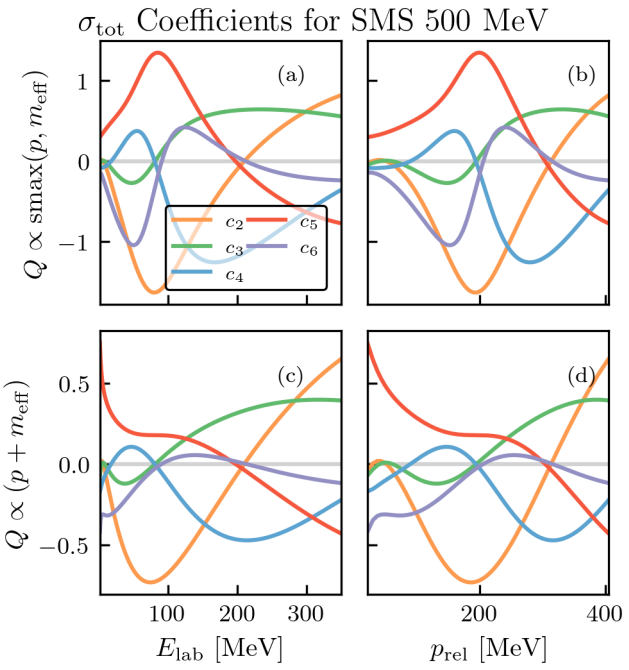



FIG. 3. Observable coefficients for the total neutron-proton cross section ($\sigma_{\rm tot}$), under various assumptions for $Q(p)$ and the input space $x$. The coefficients presented here are extracted the same way as those in Fig. 1, with $\Lambda_b = 600\,\mathrm{MeV}$ and $m_{\rm eff} = 200\,\mathrm{MeV}$.

ity and naturalness through reparametrization because of physics reasons. In this case, it remains an option to simply exclude that region from training and testing the GP (or use a nonstationary GP; see Sec. VII). For examples, see Figs. 17 and 18 (when the EFT is ill-suited to a particular domain), and Figs. 19–21 and 29–31 (where constraints distort the GP fitting process in regions where data may not be measured). The particular regime that bears mentioning is that of momenta below the pion rest mass of approximately 138 MeV, where pionless EFT is expected to describe the power counting more accurately than $\chi$EFT. Imposing the assumption of stationarity across all momenta may lead to inconsistencies and ill-fitting due to an underlying nonstationarity. Specifically, we note just such dependence of the length scale on the momentum—i.e., manifest nonstationarity in the GP—and its implications in Sec. IV.

## III. DIAGNOSTICS AND OUTPUTS

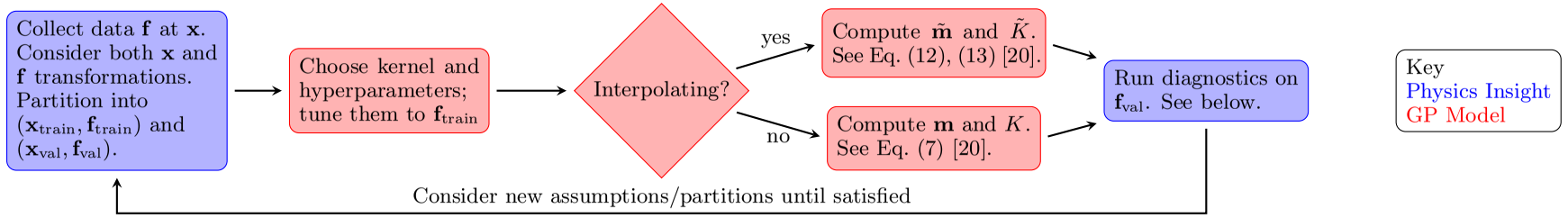
An important part of a complete Bayesian analysis is model checking [40]. Here this means assessing whether the assumptions of the BUQEYE model of $\chi$EFT convergence are validated or violated in practice. The first step is a visual examination of the order-by-order EFT coefficient curves as we saw in Sec. II: It often permits preliminary identification of flaws in the GP model for those coefficients. If one (or more than one) curve fluctuates markedly farther away from zero and/or fluctuates at a different rate than the others, that may signal a problem. Such behavior contradicts our hypothesis that all $n_c$ of the $c_n$s are drawn from a GP with a common length scale and variance. And, if the fluctuations in an EFT coefficient change significantly in size or length scale over the kinematic domain being considered, this violates the assumption of a stationary GP; i.e., that $\bar{c}^2$ and $\ell$ do not depend on the NN kinematics at which $c_n$ is evaluated. We note that GP stationarity is not a necessary assumption of the BUQEYE model (see Sec. IV), but it is taken for granted throughout the work presented in this paper because of our choice of GP kernel.

But detecting these patterns is non-trivial; humans may see patterns where there are none. Thus we consider the general statistical diagnostics for GPs that were carefully explained and tested in Ref. [41]. In Ref. [20] we adapted Ref. [41] to the particular case of EFTs. Out of the several diagnostics considered in Ref. [41] we emphasize three that we have found most illuminating in our analysis of the EFT expansion coefficients for NN potentials. These diagnostics are presented in this section, and their use is demonstrated in a publicly available Python package [42]. Our claims regarding what the different diagnostics show can be verified in that notebook. A general summary of the diagnostic workflow is given in Table II.

The default diagnostic for determining the "quality of fit" in the nuclear physics literature has usually been the reduced chi squared $\chi^2_{\rm red}$ (a.k.a. the chi squared per degree of freedom) [43]. It is often wielded as "proof" of a model's success or failure, but the theory behind its use and hence the interpretation of its value is too frequently omitted from discussion. The structure of our discussion is designed to build upon the intuition of $\chi^2_{\rm red}$ and extend it to the case of correlated errors. Some important theoretical properties of the chi squared per degree of freedom and its extensions are also provided without proof (see [41]).

Assume, as we do throughout this paper, that the EFT LECs have already been fit to data or are otherwise fixed quantities. Moreover, assume that we have computed order-by-order predictions of some observable $y_n$ at two sets of points, the $N$ training points $\mathbf{x}_{\rm train}$ and the $M$ validation points $\mathbf{x}_{\rm val}$. The quality of the generated statistical diagnostics depends not only on the choices made in Sec. II (see Table I) but also upon this so-called "train-test split." This partitioning can be revisited every time a workflow cycle completes. One sign that the user may wish to revisit the train-test split is that the coefficient curves by eye look well-suited to a GP, but the diagnostics say otherwise. In that case, the user may change the partitioning and see whether the diagnostics change (e.g., perform cross-validation). A model that proves robust under reasonable variation of the train-test split is one whose results can be taken as more credible than those of a model whose results are highly variable.

Using too few or too many training and testing points




| Diagnostic | Formula | Motivation | Success | Failure |
|---|---|---|---|---|
| Visualize the function | — | Does $\mathbf{f}_{\text{val}}$ look like a draw from a GP? What kind of GP? | $\mathbf{f}_{\text{val}}$ "looks similar" to draws from a GP | $\mathbf{f}_{\text{val}}$ "stands out" compared to GP draws |
| Mahalanobis Distance $\text{D}^2_{\text{MD}}$ | $(\mathbf{f}_{\text{val}} - \mathbf{m})^\intercal K^{-1} (\mathbf{f}_{\text{val}} - \mathbf{m})$ | Can we *quantify* how much the $\mathbf{f}_{\text{val}}$ looks like a GP? | $\text{D}^2_{\text{MD}}$ follows its theoretical distribution ($\chi^2_M$) | $\text{D}^2_{\text{MD}}$ lies too far away from the expected value of $M$ |
| Pivoted Cholesky $\mathbf{D}_{\text{PC}}$ | $G^{-1}(\mathbf{f}_{\text{val}} - \mathbf{m})$ | Can we understand why $\text{D}^2_{\text{MD}}$ is failing? | At each index, points follow standard Gaussian | Many cases (see below) |
| Credible Interval $\text{D}_{\text{CI}}(P)$ for $P \in [0,1]$ | $\frac{1}{M}\sum_{i=1}^{M} \mathbf{1}[\mathbf{f}_{\text{val},i} \in \text{CI}_i(P)]$ | Do $100P\%$ credible intervals capture data roughly $100P\%$ of the time? | Plot $\text{D}_{\text{CI}}(P)$ for $P \in [0,1]$; the curve should be within errors of $\text{D}_{\text{CI}}(P) = P$, | $\text{D}_{\text{CI}}(P)$ is far from $100P\%$, particularly for large $100P\%$ (e.g., 68% and 95%). |

| Variance | Length Scale | Observed Pattern in $\mathbf{D}_{\text{PC}}$ |
|---|---|---|
| $\sigma_{\text{est}} = \sigma_{\text{true}}$ | $\ell_{\text{est}} = \ell_{\text{true}}$ | Points are distributed as a standard Gaussian, with no pattern across index (e.g., only $\approx 5\%$ of points outside $2\sigma$ lines). |
| $\sigma_{\text{est}} = \sigma_{\text{true}}$ | $\ell_{\text{est}} > \ell_{\text{true}}$ | Points look well distributed at small index but expand to a too-large range at high index. |
| $\sigma_{\text{est}} = \sigma_{\text{true}}$ | $\ell_{\text{est}} < \ell_{\text{true}}$ | Points look well distributed at small index but shrink to a too-small range at high index. |
| $\sigma_{\text{est}} > \sigma_{\text{true}}$ | $\ell_{\text{est}} = \ell_{\text{true}}$ | Points are distributed in a too-small range at all indices. |
| $\sigma_{\text{est}} < \sigma_{\text{true}}$ | $\ell_{\text{est}} = \ell_{\text{true}}$ | Points are distributed in a too-large range at all indices. |

TABLE II. A cheatsheet for diagnostics. The interpretation of all variables and the workflow that we have found valuable is described in detail in the text.

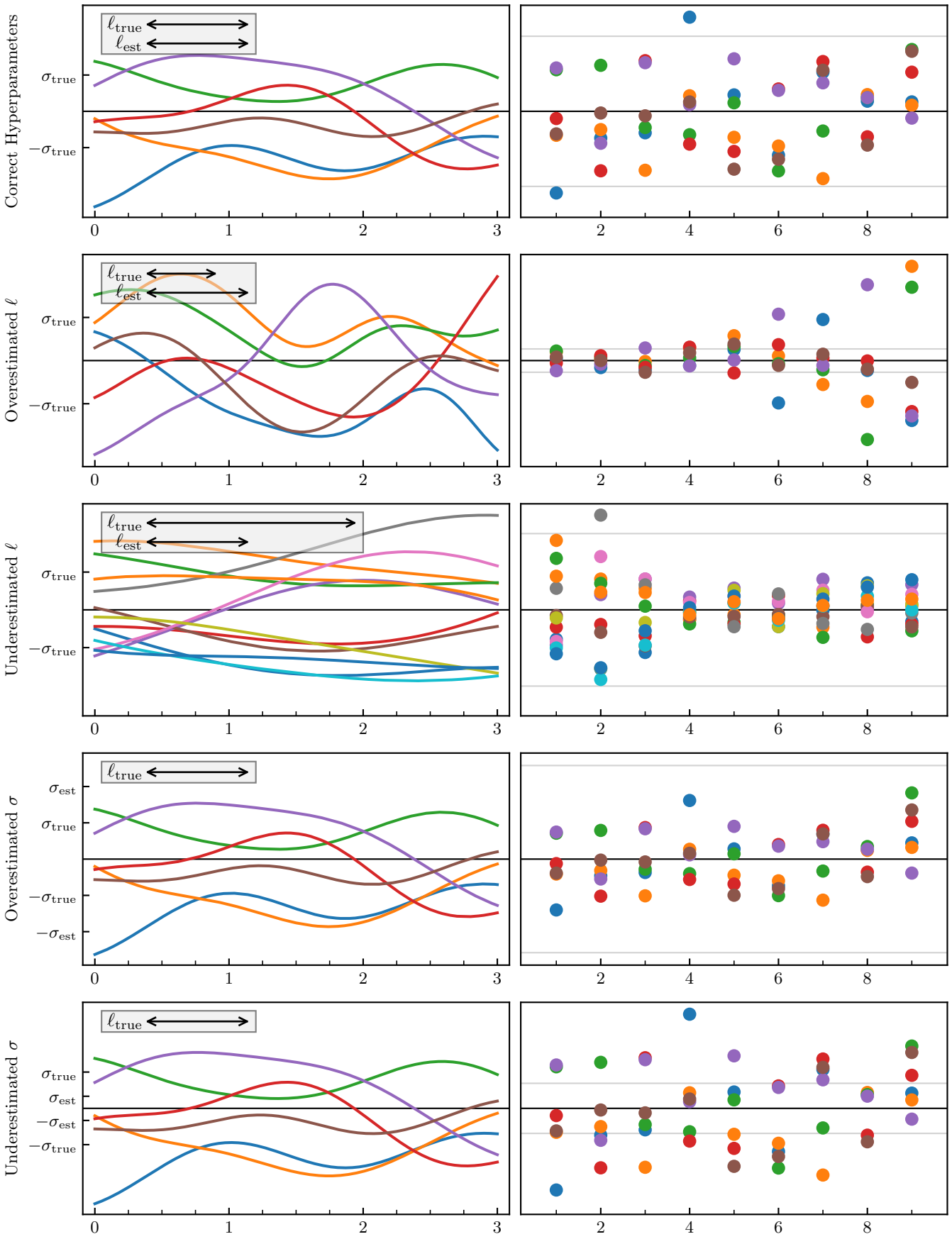



FIG. 4. Illustrations of Table II using a toy model. See text for details.

may lead to unwarranted conclusions about robustness. First, too few training or testing points can lead to a model that is undertrained and overtested or overtrained and undertested. Second, and somewhat counterintuitively, the user cannot guard against these pitfalls by using many training and testing points. Too many of either within a correlation length will lead to matrix ill-conditioning, which renders matrix inversion problematic. The diagnostics then become very sensitive to the "nugget" used to regularize the matrix inversion. Results that are sensitive to the choice of regularization of the (nearly) degenerate covariance matrix do not represent true tests of the statistical model that the $c_n$s are all drawn from a stationary GP. To strike a safe path between these two perils we used one or two training points and three or four testing points within a correlation length.

At the training points we have tuned our GP parameters $\boldsymbol{\theta}$ to its convergence pattern. At the validation points we wish to assess whether we have accurately characterized our understanding of two cases, to which we apply the workflow and diagnostics in Table II. These are:

1. The known low-order convergence pattern, by contrasting the $c_n$s with their tuned distribution. In this case, the validation data is $\mathbf{f}_{\rm val} \equiv \mathbf{c}_n$ for some $n \leq k$ and it is to be compared to the multivariate Gaussian with mean $\mathbf{m} = \mathbf{0}$ and covariance $K = \bar{c}^2 R$. This analysis helps answer the question of whether the EFT converges as we expect. The "visualize the function" diagnostic asks: "Do coefficient curves look like they are all drawn from the same GP?" (In)consistency can be tricky to cleanly assess at this stage; for example, all curves may look to have similar length scale and variance (apparently consistent), or one (or more) curves can be seen to have a distinct variance or length scale (apparently inconsistent). But beware that fluctuations happen, so what looks dissimilar to humans may be the randomness.

2. The unknown truncation error, by contrasting $y_k$ with experimental data $y_{\rm exp}$, including all sources of uncertainty. In this case, the validation data is $\mathbf{f}_{\rm val} \equiv \mathbf{y}_{\rm exp}$ and it is to be compared to the multivariate Gaussian with mean $\mathbf{m} = \mathbf{y}_k$ and covariance $K = \Sigma_{th} + \Sigma_{exp}$. This analysis helps answer the question of whether the EFT predictions are consistent with experiment when all errors are accounted for.

With the notation for these cases defined, we can now step through the other diagnostics in Table II.

### A. Mahalanobis distance

The chi-squared statistic is an intuitive diagnostic to consider first, and can help introduce the idea of a reference distribution. It is defined by

$$\chi^2 = \sum_{i=1}^{M} \frac{(f_{{\rm val},i} - m_i)^2}{\sigma_i^2}. \tag{13}$$

If the $\mathbf{f}_{\rm val}$ were really drawn from a Gaussian with mean $\mathbf{m}$ and uncorrelated noise $\sigma_i$, then this aptly named quantity would follow a $\chi^2_\nu$ distribution with $\nu \equiv M$ degrees of freedom. It is important to stress that Eq. (13) has a *distribution*; that is, even if we had correctly estimated $\mathbf{f}_{\rm val}$, it could be above or below the "gold standard" value, but it should not be too far away. The mean of the $\chi^2_\nu$ distribution is $\nu$. To naïve $\chi^2_{\rm red}$ apologists it then follows, as night follows day, that $\chi^2_{\rm red} = \chi^2/\nu$ *should* be 1. This claim should be regarded with skepticism unless error bands are provided (and with extreme skepticism if one is using the $\chi^2_{\rm red}$ as a goodness-of-fit criterion for a non-linear model, see [43]). That is, the reference distribution $\chi^2_\nu$, both its mean and its uncertainties, can inform us what reasonable $\chi^2$ diagnostic sizes look like.

For the truncation error model described in Sec. II, it is clear that the theoretical errors become correlated. Thus, the vanilla chi-squared statistic no longer applies. However, one can compute the (squared) Mahalanobis distance

$$\mathrm{D}^2_{\rm MD}(\mathbf{f}_{\rm val}) = (\mathbf{f}_{\rm val} - \mathbf{m})^\intercal K^{-1} (\mathbf{f}_{\rm val} - \mathbf{m}), \tag{14}$$

of which the chi-squared is merely a special case when the errors are uncorrelated. In calculating this distance, the covariance matrix of the GP at the validation points, $K$, plays the role of a metric tensor. Interestingly, if $\mathbf{f}_{\rm val}$ were really drawn from a multivariate Gaussian with mean $\mathbf{m}$ and covariance $K$, then $\mathrm{D}^2_{\rm MD}$ still follows a $\chi^2_\nu$ distribution with $\nu = M$ degrees of freedom.

The $\mathrm{D}^2_{\rm MD}$ combined with its reference distribution can quantitatively tell us whether, for example, our observable coefficients $c_n$ follow a GP as we hypothesize. We evaluate $\mathrm{D}^2_{\rm MD}(\mathbf{f}_{\rm val})$ separately at each order (i.e., for each $c_n$ curve) for which we have a $\chi$EFT calculation of NN scattering; that is, we use the $\chi$EFT coefficients at the kinematics defined by the validation points to compute the Mahalanobis distance of those coefficients from the mean curve——which is taken to be zero in our model of EFT coefficients. This allows us to see if all $n_c$ of the $\mathrm{D}^2_{\rm MD}(\mathbf{c}_n)$ values lie within a reasonable range (68% or 95%) of the reference distribution (which shows that the curves are consistent with our model) or whether there are outliers (which shows the opposite). If one or more of the EFT coefficients correspond to a GP fit that is statistically too good then $\mathrm{D}^2_{\rm MD}$ for those $c_n$s will be markedly less than the number of degrees of freedom because errors in the GP are overestimated. In contrast, a $\mathrm{D}^2_{\rm MD}$ that is large compared to the number of degrees of freedom means the errors in the GP do not encompass the validation points in a statistically correct way.

### B. Pivoted Cholesky decomposition

Instead of relying on a one-number summary, such as the chi squared statistic, one could instead consider the weighted residual $(f_{\mathrm{val},i} - m_i)^2/\sigma_i^2$ at each point. Such a residual vector contains much more information than its sum, and permits one to inquire where exactly the theory is failing. If $\mathbf{m}$ and $\sigma$ are correct, then there should be no pattern across the indices of the residual vector, and the reference distribution for each index is itself a standard Gaussian $\mathcal{N}(0,1)$.

The correlated analog requires a standard deviation matrix $K = GG^\intercal$ from which one can compute

$$\mathbf{D}_G = G^{-1}(\mathbf{f}_{\mathrm{val}} - \mathbf{m}). \tag{15}$$

The $G$ matrix is not unique, but we choose it to be the pivoted Cholesky decomposition [41], and call $\mathbf{D}_{\mathrm{PC}}$ the pivoted Cholesky (PC) diagnostic. Each index of this vector still corresponds to one particular validation point, but the indices have been pivoted such that the first index has the largest variance, the second has the largest variance after one has conditioned on the first validation point, and so on [41]. Again, the reference distribution at each index is a standard Gaussian, but this diagnostic can fail in illuminating ways. In essence, misestimates of the variance $\bar{c}^2$ appear at all indices, and misestimates of the correlation structure show up at large index.

Table II lists five possible patterns for this diagnostic, which are illustrated in Fig. 4. Interpretations of the corresponding plots of the $c_n$ and PC diagnostics include (with length scale $\ell$, standard deviation $\sigma$, GP-estimated quantities "est," and actual underlying quantities "true"):

1. Correct: At a given index, coefficient values at different orders are Gaussian-distributed, with the same variance exhibited at all indices (see subplot (a) in Fig. 4).

2. $\ell_{\mathrm{est}} > \ell_{\mathrm{true}}$: EFT coefficients associated with different orders are correctly distributed at small index but their variance gets noticeably larger as the index increases (see subplot (b) in Fig. 4), a phenomenon known as "trumpeting."

3. $\ell_{\mathrm{est}} < \ell_{\mathrm{true}}$: EFT coefficients associated with different orders are correctly distributed at small index but their variance gets noticeably smaller as the index increases (see subplot (c) in Fig. 4), a phenomenon known as "funneling."

4. $\sigma_{\mathrm{est}} > \sigma_{\mathrm{true}}$: EFT coefficients at different orders exhibit scatter that is smaller than the estimated variance, and this happens across all validation points (see subplot (d) in Fig. 4).

5. $\sigma_{\mathrm{est}} < \sigma_{\mathrm{true}}$: EFT coefficients at different orders exhibit scatter that is larger than the estimated variance, and this happens for all validation points (see subplot (e) in Fig. 4).

In our EFT application we examine $\mathbf{D}_{\mathrm{PC}}$ for each order in the EFT for which we have coefficient data—as we did with $\mathrm{D}^2_{\mathrm{MD}}$—since it may be that the variance or length-scale problems we are trying to diagnose show up at some orders but not others.

### C. Credible interval diagnostic

We seek credible intervals that are accurate representations of our uncertainty. For a given set of $c_n$s, the GP is trained on all the coefficients at their training points, the underlying distribution is calculated from the fitted GP. We assess the accuracy of our credible intervals by comparing the coefficients' distributions at their validation (testing) points to this underlying distribution. If our model for the uncertainty is accurate and we construct a $100P\%$ credible interval for $P \in [0,1]$, it should approximately contain $100P\%$ of the validation data. Our final diagnostic implements this idea, and is known as the credible interval diagnostic $\mathrm{D}_{\mathrm{CI}}$, the empirical coverage probability plot, or, more colloquially, the weather plot [4]. One can check this at a single value of $P$ or all values between $[0,1]$. We plot $\mathrm{D}_{\mathrm{CI}}$ for all $P$, denoted $\mathrm{D}_{\mathrm{CI}}(P)$, and create its reference distribution via sampling: This determines how far away $\mathrm{D}_{\mathrm{CI}}(P)$ can be from $P$ before the diagnostic signals statistical inconsistency [20].

Examination of a plot of $\mathrm{D}_{\mathrm{CI}}(P)$ vs. $P$ will therefore quickly reveal whether the estimated GP variance is too large (entailing too-large error bands), in which case $\mathrm{D}_{\mathrm{CI}}(P)$ grows faster than $P$, or too small, leading to $\mathrm{D}_{\mathrm{CI}}(P)$ consistently smaller than $P$. An example of the former case is shown in Fig. 5a, in which too small a value of $m_{\mathrm{eff}}$ leads to too large a GP variance, a situation that is visibly rectified in Fig. 5b when the value of $m_{\mathrm{eff}}$ is increased.

The $\mathrm{D}_{\mathrm{CI}}$ diagnostic is an internal check on the self-consistency of our model; it does not depend on experimental data. We can, however, also compare EFT predictions with full uncertainty quantification to data [17, 39, 45]. Such a comparison is made in the next section.

### D. Output: Statistically rigorous EFT error bands for observables

If the GP model passes all the other diagnostic tests, then we can use Eq. (4) to form statistically consistent EFT truncation uncertainty bands for the observable(s) from which the $c_n$s were extracted. We typically do this using the MAP values of $\Lambda_b$ and $m_{\mathrm{eff}}$. The resulting error bands then include not only the EFT truncation uncertainty, but also the GP uncertainties from interpolating

[4] The term "weather plot" is inspired by Ref. [44], which explains in more detail the origin of this nickname.

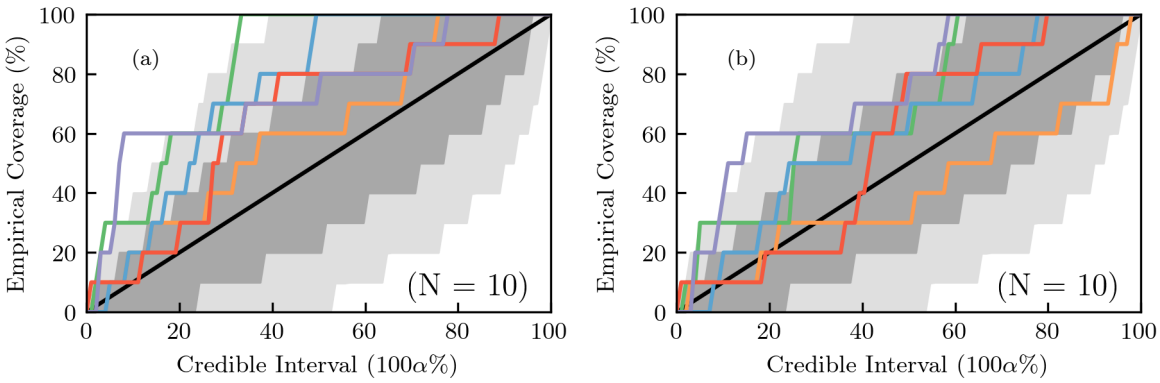



FIG. 5. Plots of the credible intervals ("weather plots") corresponding to the coefficients of the total cross section in Fig. 1d (a) and Fig. 3d (b). The concentration of curves above and to the left of the black midline in the lefthand figure implies that the GP variance is being overestimated so the credible intervals are too large. The curves on the righthand side track the midline better: The empirical coverage falls within the shaded region, so the error bands better capture the shift at the next EFT order.

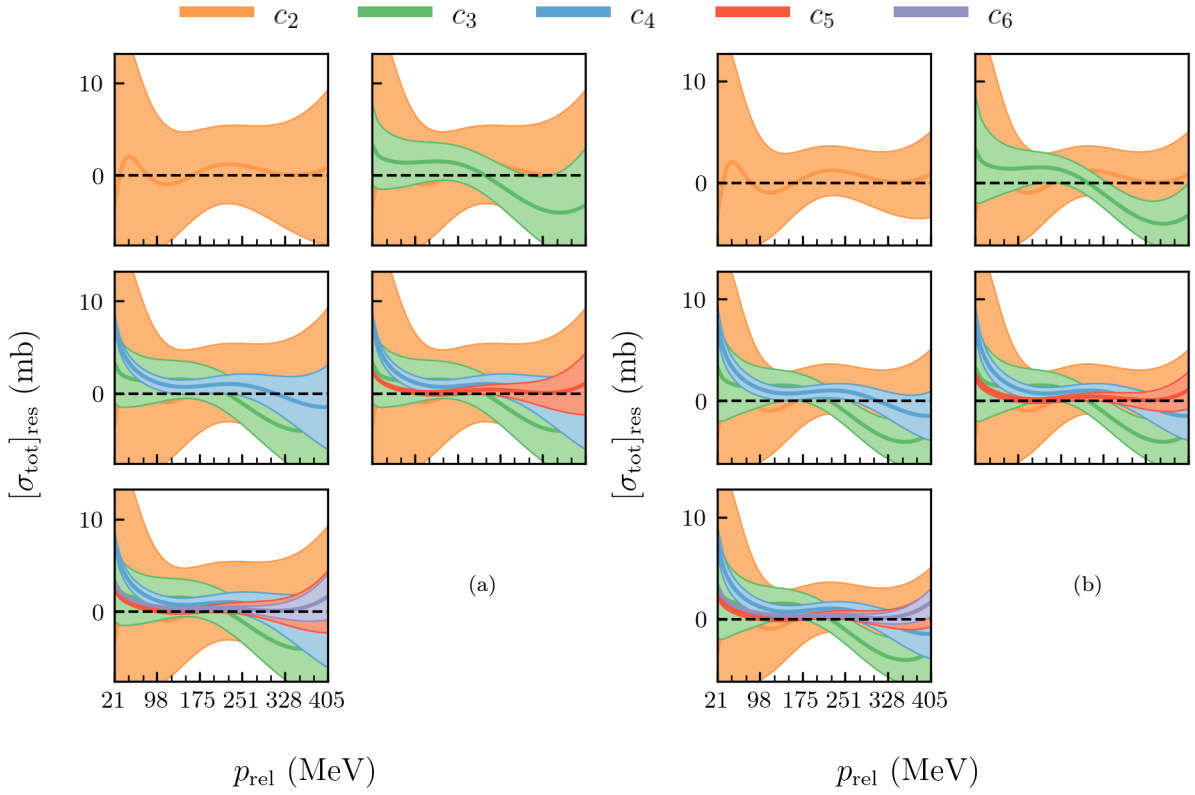



FIG. 6. Plots of the residuals and truncation error corresponding to the coefficients of the total cross section in Fig. 1d (a) and Fig. 3d (b). Here, the differences between the experimental and predicted value of the total cross section at each order (the solid colored lines) with their associated same-color 68% error bands, are plotted against zero (the dashed black line). This is useful for assessing whether the error model is converging properly (e.g., falling within the predicted 68% error bands 68% of the time) and where in the input space the predictions agree most and least with experiment. Here, one can observe improvement in agreement between the experimental and theoretical values for the observable as the value of $m_{\rm eff}$ is changed from left to right, but at higher orders one sees degraded performance at low momentum, which accords with our visual assessment of the corresponding coefficient plots.

between the training points. They do not include the uncertainties in the GP hyperparameters since we use point estimates for the length scales and $\bar{c}^2$.

For an example of truncation-error plots that builds

on the last section's comparison of Figs. 1d and 3d, see Fig. 6. This shows the residuals between the true and predicted values for $\sigma_{\rm tot}$ at each of the orders under test. The difference between the two sets of plots arises from choosing $m_{\rm eff} = 138\,{\rm MeV}$ (6a) versus $m_{\rm eff} = 200\,{\rm MeV}$ (6b), which results in different error bands. The improvement occasioned by the change in $m_{\rm eff}$'s value is visible in the superior agreement between the true and predicted values in the right-hand plot. That the low-momentum region is where the greatest discrepancy can be found flags that regime as an area where our statistical model may be less well-suited.

### E. Output: Posterior probability distributions of $\Lambda_b$ and $m_{\rm eff}$

One circumstance in which the statistical diagnostics immediately reveal a problem is if the EFT expansion parameter has been misestimated. In that case the variance of the coefficients $c_n$ will either grow or shrink with EFT order $n$: grow if the expansion parameter has been chosen too small, and shrink if it has been chosen too large. In either case the $D_{\rm CI}$ plot shows a failure to correctly estimate Bayesian credible intervals, with that failure increasing in severity at higher EFT orders. Of course, this kind of systematic trend in the $c_n$s can sometimes be seen when they are plotted together on the same scale. This indication of trouble from the $c_n$ curves can be quantified, in this case via a formula that estimates the best value of $\Lambda_b$ [17, 20]. The method employed in Refs. [17] was pointwise — that is, it treated the estimate of the uncertainty at each testing point as independent of that at other testing points — whereas the newer methods showcased in this paper are curvewise — that is, they account for correlations among testing points at a given order by means of a length scale that characterizes a GP.

The best values of $\Lambda_b$ and $m_{\rm eff}$ are the ones that "right-size" the coefficients, i.e., ensure that the coefficients appear to be drawn from the same distribution. The requirement that they do not show a systematic size trend thus translates into a probability distribution function (pdf) for the EFT breakdown momentum and effective soft scale (see the Appendices of Ref. [20]).

To calculate this pdf, we first set meshes for each of the $n_r$ hyperparameters. In our case, $n_r = 4$: $\Lambda_b$, $m_{\rm eff}$, the length scale in the lab-energy input space $\ell_E$, and the length scale in the scattering-angle input space $\ell_\theta$. Then, we form a $n_r$-dimensional mesh from the Cartesian product of these one-dimensional meshes. We have not listed the marginal variance $\bar{c}^2$ as one of the random variables that forms the mesh, because Gaussian processes can be marginalized over their variance analytically. This yields a statistical object formed out of the Student $t$-distribution instead of one formed out of a Gaussian (normal) distribution—a TP rather than a GP. So, at each point in this $n_r$-dimensional mesh we calculate the log-likelihood that the $\chi$EFT coefficients $c_n$ for a given observable are described by a TP corresponding to the hyperparameters associated with that location in the mesh. We add the log-priors to the log-likelihood to find the log-posterior, and exponentiate & normalize it to find the posterior pdf on the mesh. (When we learn from more than one observable at once we assume they are independent, by summing the log-likelihoods before combining the overall log-likelihood with the log-priors.) Lastly, we marginalize over the length scales, since $\Lambda_b$ and $m_{\rm eff}$ are physical parameters across observables and length scales are observable-specific. From the joint $\Lambda_b$-$m_{\rm eff}$ posterior, we can extract MAP values for each of these two random variables, calculate their correlation coefficient, and also marginalize to find their one-dimensional posterior pdfs, including the respective means and variances of $\Lambda_b$ and $m_{\rm eff}$.

A key question is then whether the $\Lambda_b$ and $m_{\rm eff}$ values initially used to form the $c_n$s from the observable coefficients are consistent with this posterior. If not, the analysis has to be repeated with values that are — this would typically be the MAP value of the two-dimensional pdf ${\rm pr}(\Lambda_b, m_{\rm eff}|\mathbf{y}_k, I)$.

The results of these calculations are discussed in detail in the next section, Sec. IV. Note that we only use TPs for computing the posterior pdfs of $\Lambda_b$ and $m_{\rm eff}$, since for those physical parameters we want to marginalize over $\bar{c}^2$. The coefficients $c_n$ of the $\chi$EFT expansion are still fit with a GP whenever we generate the diagnostics that tell us whether the EFT coefficients conform to the BUQEYE statistical model or not.

## IV. GP STATIONARITY AND $\Lambda_b$ AND $m_{\rm eff}$

Previous work placed strong (even delta-function) priors on the values of $\Lambda_b$ and $m_{\rm eff}$, treating them as point values. Continuing to treat them this way would be acceptable as long as we had ironclad intuition on a range in which those values were likely to fall and the results of our analysis were not sensitive to them, but our intuition is not strong enough and our outcomes not insensitive enough to justify that approach. Indeed, we have already seen in Fig. 5 the improvement made in the stationarity and naturalness of the $\sigma_{\rm tot}$ coefficients when the value of $m_{\rm eff}$ is changed from 138 to 200 MeV from Fig. 1 to Fig. 3. Thus, more care is needed in specifying the values of those parameters for the purposes of generating graphical and statistical diagnostics.

We begin with an extraction of the expansion parameter $Q$ that uses separate and independent one-dimensional TPs to model the EFT coefficients at several different energies. In the terminology deployed above, this approach is pointwise in $x_E$ and curvewise in $x_\theta$. We considered the EFT coefficients of the differential cross section and the five spin observables $D$, $A_{xx}$, $A_{yy}$, $A$, and $A_y$ at different fixed values of $p_{\rm rel}$ from 25 to 400 MeV in increments of 25 MeV and employed the procedure discussed in Sec. III E (*mutatis mutandis* since we only have

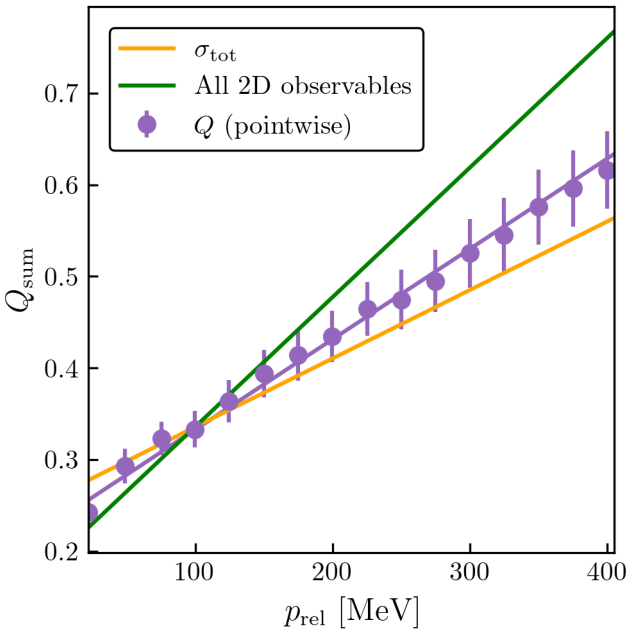



FIG. 7. The MAP values of the dimensionless expansion parameter $Q$ extracted using the procedure outlined in Sec. III E and corresponding $2\sigma$ error bars are shown in purple as a function of the relative momentum $p_{\rm rel}$. The corresponding linear fit, which is used to extract $\Lambda_b^\star$ and $m_{\rm eff}^\star$ in Eq. (16), is also shown in purple. Also shown are: the orange line $Q = Q_{\rm sum}(p = p_{\rm rel}, \Lambda_b, m_{\rm eff})$, with $\Lambda_b$ and $m_{\rm eff}$ MAP values extracted from posterior pdfs generated from $\sigma_{\rm tot}$ (see the ninth row of Table III), and the green line $Q = Q_{\rm sum}(p = p_{\rm rel}, \Lambda_b', m_{\rm eff}')$, with $\Lambda_b$ and $m_{\rm eff}$ values obtained as per the first row of Table III [see Eq. (17)]. All information shown in this figure comes from observable predictions up to and including the highest order under consideration, N$^4$LO+.

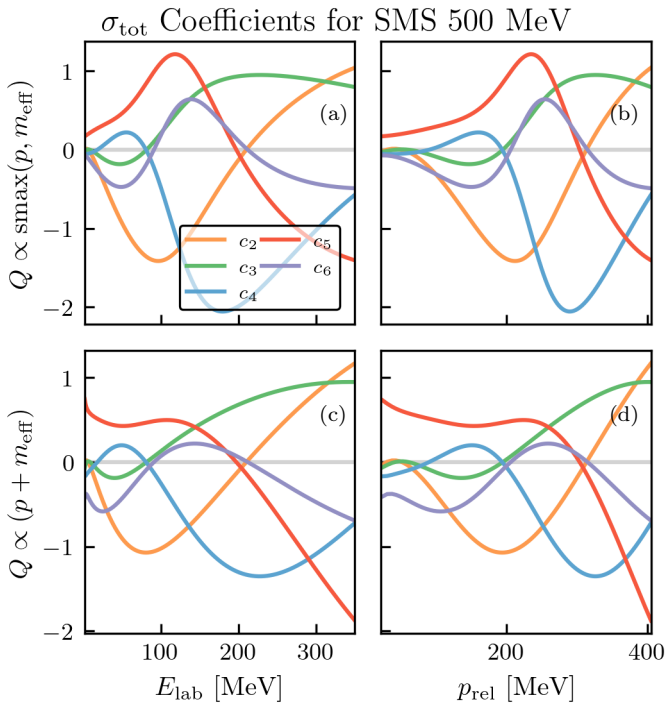



FIG. 8. Observable coefficients for the total neutron-proton cross section ($\sigma_{\rm tot}$), under various assumptions for $Q(p)$ and the input space $x$. The coefficients presented here are extracted the same way as those in Fig. 3 except that the values of $\Lambda_b$ and $m_{\rm eff}$ are the MAP values extracted from the corresponding joint posterior probability distributions (see Table III). Specifically, subplots (a) and (b) are plotted using $\Lambda_b$ and $m_{\rm eff}$ values from the tenth row of Table III and subplots (c) and (d) using values from the ninth row.

the angular length scale) to extract a posterior pdf for $Q$ at each $p_{\rm rel}$. Figure 7 shows the $2\sigma$ confidence intervals for $Q(p_{\rm rel})$ that result from these 16 one-dimensional analyses. There is a clear linear dependence on $p_{\rm rel}$ that strongly supports the $Q_{\rm sum}$ parametrization of the EFT expansion parameter. The straight-line fit to the $Q$ data is shown in purple in the figure and yields:[5]

$$\Lambda_b^\star = 780 \pm 20\,\text{MeV} \quad \text{and} \quad m_{\rm eff}^\star = 240 \pm 10\,\text{MeV}. \quad (16)$$

How do these values relate to those we get from the total cross section, which was not part of the set of observables used to extract $Q(p_{\rm rel})$? The intervals for $\Lambda_b$ and $m_{\rm eff}$ obtained when a TP is used to analyze the EFT convergence of only $\sigma_{\rm tot}$ are given in the ninth and tenth rows of Table III, where the ninth (tenth) row uses the $Q_{\rm sum}$ ($Q_{\rm max}$) parametrization of the expansion parameter. Other rows show those intervals (also extracted via TP) for different analysis choices and are discussed below. The intervals from the $\sigma_{\rm tot}$ analysis are not consistent with the numbers in Eq. (16), but the resulting plot of $Q$ that the $\sigma_{\rm tot}$ values yield (plotted as an orange line in Fig. 7) is somewhat similar (within $2\sigma$ error bands) to those yielded by the numbers in Eq. (16).

Furthermore, these extracted values for the breakdown and soft scales correspond to curves that meet our criteria for stationarity and naturalness. In Fig. 8 we replot the $\sigma_{\rm tot}$ coefficients from Figs. 1 and 3 using the MAP values for $\Lambda_b$ and $m_{\rm eff}$ for both $Q$ parametrizations. Figures 8a and 8b show the coefficients using the values in the $Q_{\rm sum}$ $\sigma_{\rm tot}$ (ninth) row of Table III and Figs. 8c and 8d those obtained with the values in the $Q_{\rm smax}$ $\sigma_{\rm tot}$ (tenth) row. The result is coefficient functions that are stationary and natural to the unaided eye across momenta.

However, the values extracted from $\sigma_{\rm tot}$ data alone are not consistent with Epelbaum's suggestion of $\Lambda_b \approx$ 650–700 MeV and $m_{\rm eff} \approx$ 200–225 MeV. These breakdown and soft scale ranges were obtained from empirical coverage plots that optimize $m_{\rm eff}$ for the highest-order predictions of $\sigma_{\rm tot}$ alone [46].

What does a full two-dimensional analysis of the EFT coefficient curves reveal? The posterior pdfs for

[5] Unless otherwise noted, these intervals and others discussed in this section are 68% credibility intervals, which corresponds to $1\sigma$.

Extracted Values of $\Lambda_b$ and $m_{\rm eff}$ in MeV for Different Parametrizations

| $Q$ | $p$ | $x_E$ | $x_\theta$ | $\Lambda_b$ | $m_{\rm eff}$ | Observable(s) | Comments |
|---|---|---|---|---|---|---|---|
| $Q_{\rm sum}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | $-\cos(\theta)$ | $570 \pm 10$ | $138 \pm 3$ | All 2D obs. | Figs. 7, 11–13, 15, 17–20, 22–24, 26b, 27b, 29–30, 32–33. |
| $Q_{\rm smax}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | $-\cos(\theta)$ | $378 \pm 5$ | $106 \pm 0$ | All 2D obs. | Fig. 10. |
| $Q_{\rm sum}$ | $p_{\rm rel}$ | $E_{\rm lab}$ | $-\cos(\theta)$ | $610 \pm 10$ | $186 \pm 4$ | All 2D obs. | Fig. 16, 27a. |
| $Q_{\rm smax}$ | $p_{\rm rel}$ | $E_{\rm lab}$ | $-\cos(\theta)$ | $459 \pm 6$ | $155 \pm 1$ | All 2D obs. | |
| $Q_{\rm sum}$ | $p_{\rm smax}(p_{\rm rel}, q_{\rm CM})$ | $p_{\rm rel}$ | $-\cos(\theta)$ | $660 \pm 10$ | $172 \pm 4$ | All 2D obs. | Figs. 26a, 28. |
| $Q_{\rm sum}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | $q_{\rm CM}$ | $650 \pm 10$ | $184 \pm 5$ | All 2D obs. | Fig. 14. |
| $Q_{\rm sum}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | $\theta$ | $590 \pm 10$ | $144 \pm 2$ | All 2D obs. | Figs. 21, 31. |
| $Q_{\rm sum}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | $-\cos(\theta)$ | $530 \pm 10$ | $120. \pm 3$ | All 2D obs. | $c_6$ omitted. Fig. 25. |
| $Q_{\rm sum}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | | $990 \pm 90$ | $350 \pm 40$ | $\sigma_{\rm tot}$ | Figs. 7, 8c–8d. |
| $Q_{\rm smax}$ | $p_{\rm rel}$ | $p_{\rm rel}$ | | $670 \pm 70$ | $250 \pm 40$ | $\sigma_{\rm tot}$ | Figs. 8a–8b. |

TABLE III. This table includes information on the posteriors (as calculated with a TP per Sec. III E) of the breakdown scale $\Lambda_b$ and soft scale $m_{\rm eff}$ under different analysis choices. The first through fourth columns list the choices of parametrization for $Q$, $p$, the lab-energy input space $x_E$, and the scattering-angle input space $x_\theta$; the fifth and sixth columns list the mean values and standard deviations for the fully marginalized $\Lambda_b$ and $m_{\rm eff}$ posterior pdfs; the seventh column shows the observable(s) from which the values are derived; and the eighth column lists the figures in this paper in which these MAP values are used. For all calculations involving the 2D observables, the $\Lambda_b$ mesh ranges from 200 to 900 MeV, the $m_{\rm eff}$ mesh from 1 to 350 MeV, the $\ell_E$ mesh from 1 to 150 MeV, and the $\ell_\theta$ mesh from 0.01 to 2 times the total length of the scattering-angle input space. For all calculations involving $\sigma_{\rm tot}$, the length scale meshes are the same as for calculations involving the 2D observables but the $\Lambda_b$ mesh is evenly spaced from 450 to 1150 MeV and the $m_{\rm eff}$ mesh evenly spaced from 100 to 450 MeV for the calculations done in the ninth row. Training points are located at {1, 12, 33, 65, 108, 161, 225, 300} MeV lab energy and {41, 60, 76, 90, 104, 120, 139}° scattering angle. Uniform log-priors are placed on the length scales over all positive values, a uniform log-prior is placed on $\Lambda_b$ from 200 to 900 MeV, and a uniform log-prior is placed on $m_{\rm eff}$ from 1 to 350 MeV.

the breakdown scale $\Lambda_b$ and soft scale $m_{\rm eff}$ obtained from data across the full two-dimensional input spaces $(x_E, x_\theta) = (p_{\rm rel}, -\cos(\theta))$ (the total cross section $\sigma_{\rm tot}$, for which we have only one-dimensional data, is omitted from the set of considered observables) are summarized in the first row of Table III. The intervals are:

$$\Lambda_b' = 570 \pm 10\,\text{MeV} \quad \text{and} \quad m_{\rm eff}' = 138 \pm 3\,\text{MeV}. \quad (17)$$

In this first row we have used $Q_{\rm sum}$, as well as making several other analysis choices that the next section will show improve the statistical consistency of our GP description. Other analysis choices are represented in the second to eighth rows of the table. All but one have in common that, when we learn from all the 2D observables at once, $\Lambda_b$'s 68% interval lies between 450 and 670 MeV and $m_{\rm eff}$'s between 115 and 190 MeV.

The stark differences between $(\Lambda_b^\star, m_{\rm eff}^\star)$ and $(\Lambda_b', m_{\rm eff}')$ immediately prompt the question: "Why is there so great an inconsistency?" The difference arises from the imposition of a stationary TP on what turns out to be a nonstationary set of coefficient curves. In Fig. 7, the $Q$ posterior pdfs (from which $\Lambda_b^\star$ and $m_{\rm eff}^\star$ are extracted) are calculated using data at different fixed momenta, with the resulting coefficients from which the $Q$ distributions are extracted depending on $x_\theta$ alone. Thus, while the TP used to extract the $Q$ posterior distribution at each fixed $p_{\rm rel}$ "knows" about correlations in length scale and variance across $x_\theta$, it doesn't "know" about correlations across $x_E$ (i.e., between fixed $p_{\rm rel}$). Only the TP used to calculate $\Lambda_b$ and $m_{\rm eff}$ posteriors from the two-dimensional data on EFT coefficients takes into account their correlations across $x_E$ *and* $x_\theta$.

In principle, such a two-dimensional analysis is superior, since it leverages information across energies in a way that accounts for correlations. But it is only superior if those correlations are well modeled. Recall that we assumed a stationary GP, in which the correlations between coefficients across angles persist to the same extent irrespective of the value of the momentum. However, in the lower-energy regions of $x_E = p_{\rm rel}$, coefficients in $x_\theta$ are best fit by GPs with longer length scales, while in higher-energy regions coefficients in $x_\theta$ are best fit by GPs with shorter length scales. An explanation for $\ell_\theta$'s dependence on $x_E$ is that, semiclassically, the highest partial-wave contribution $L_{\rm max}$ accessible by a state depends upon the energy of that state; thus, at low energy,

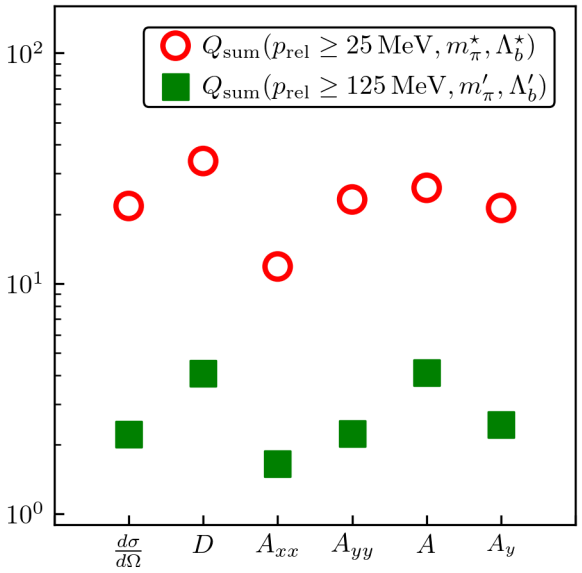



FIG. 9. Plot of the ratio of the greatest to lowest value of $\bar{c}^2$ of the coefficients at each of the fixed $p_{\rm rel}$ from Fig. 7. Red markers correspond to those generated with the values in Eq. (16) and green to those generated with the values from Eq. (17), while circle markers correspond to those generated taking into account all fixed $p_{\rm rel}$ and square to those generated taking into account only fixed $p_{\rm rel} \geq 125\,{\rm MeV}$. They are shown for the combination of all 2D observables, namely the differential cross section and the spin observables $D$, $A_{xx}$, $A_{yy}$, $A$, and $A_y$. The clear conclusion is that the results in green are more conducive to an assumption of naturalness than those in red, since the former are generally closer to 1 than the latter.

$L_{\rm max}$ too is low. The lower the quantum number $L$ of the partial wave, the less structure it has and the more slowly it changes, which renders it better characterized by a GP with a long length scale. Thus, the length scale in the angle-dependent input space $x_\theta$ should depend on the energy at which the observable is fixed and specifically be long when that energy is low. These observations and this physics argument imply that correlations in angle are not, in fact, independent of energy and a nonstationary model for GP coefficients would be a better choice for our 2D analysis.

A gauge of rough stationarity in $\bar{c}^2$, which is the GP variance and thus the characteristic size of the coefficients, can be found by taking the ratio between the greatest and least $\bar{c}^2$ values of a fitted GP at different training points across $x = (x_E, x_\theta)$. A sign that the coefficients are right-sized across $x$—i.e., stationary in $\bar{c}^2$—is that the values of that ratio hover around 1, with values much greater signaling nonstationarity in coefficient size. To assess nonstationarity in $x_E$, we compare $\bar{c}^2$ at each of the fixed $p_{\rm rel}$ from the analysis that produced Fig. 7. Specifically, we run this analysis for the differential cross section and the five spin observables. For each, we take the ratio, across the GPs fitted at each fixed momentum, of the greatest variance to the least.

We plot these values in Fig. 9, where the red circles correspond to values of this ratio extracted using $\Lambda_b^\star$ and $m_{\rm eff}^\star$ from Eq. (16). These exhibit a high degree of nonstationarity. One might think that, since the nonstationarity in $\ell_\theta$ is most notable for relatively low $p_{\rm rel}$, we can reevaluate the ratio but with the results from $p_{\rm rel} \leq 100\,{\rm MeV}$ omitted from the assessment (this is also the momentum region where pionless EFT applies to NN scattering). But even then, the ratio between the maximum and minimum is still too high. However, if we perform the reevaluation using $\Lambda_b'$ and $m_{\rm eff}'$ from Eq. (17), which were determined in an analysis that took into account correlations in $\bar{c}^2$ across $x$ *and* we omit $p_{\rm rel} \leq 100\,{\rm MeV}$ when computing the ratio, then the results correspond to the green square markers in Fig. 9. These coefficients are reasonably stationary.

This test is important for developing a preference of values for $\Lambda_b$ and $m_{\rm eff}$ and determining the physical regime in which our model is most reliable, but the true test is the one undertaken in the next section, where coefficients are extracted using different values of the breakdown and soft scales as well as different choices of parametrization and their consistency with the BUQEYE model assessed with the aid of graphical and statistical diagnostics. In that section, we will show that the choices of parametrization from the first row of Table III [namely, $Q(p) = Q_{\rm sum}(p_{\rm rel})$ and $(x_E, x_\theta) = (p_{\rm rel}, -\cos(\theta))$] that give rise to $\Lambda_b'$ and $m_{\rm eff}'$ of Eq. (17) are generally superior. When comparing this with other analysis choices in each case we adopt the corresponding MAP values of $\Lambda_b$ and $m_{\rm eff}$, since those values right-size coefficients across momenta and angle. We also decline to look at coefficients at fixed $p_{\rm rel} \leq 100\,{\rm MeV}$ due to the strength of nonstationarity's effects there. We will investigate this nonstationarity further in future work that will also address other potentials.

## V. ASSESSING THE BUQEYE MODEL FOR CHIRAL EFT

Now we can use the diagnostics from Sec. III to make a wide-ranging assessment of the choices of parametrization in Sec. II B, which lead to the favored choices summarized in Table I. Our goal here is to answer the question, "Are the dimensionless coefficients extracted according to Eq. (2) consistent with the hypothesis that they are random draws from the same GP?" If they are, that implies a pattern for the coefficients that can be inductively generalized to obtain a (correlated) truncation error for the EFT series. As we have stressed, we advocate a combination of graphical and statistical diagnostics that quantify and standardize the assessment.

In Secs. V A-V D, we present a series of case studies applying the workflow of Table II. Starting from uninformed choices for input parametrizations, we iterate different choices of $y_{\rm ref}$, $Q(p)$, and $x$, looking for general

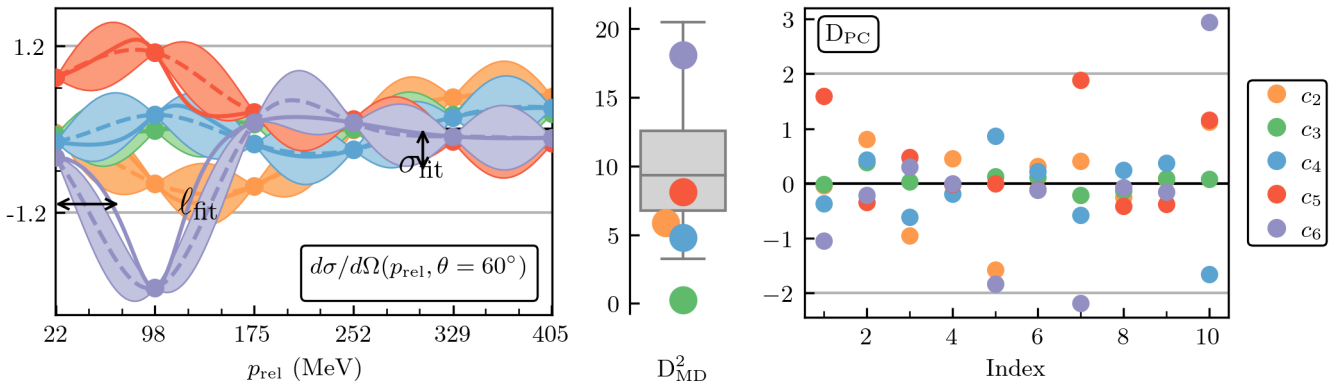



FIG. 10. Diagnostics for the differential cross section at $\theta = 60^\circ$. Here, the coefficients are plotted with $x_E = p_{\rm rel}$ and $Q = Q_{\rm smax}(p = p_{\rm rel}, m_{\rm eff} = 106\,{\rm MeV}, \Lambda_b = 378\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. Note the coefficients' generally higher variance and shorter length scale in the left half of the input space by comparison to the right half (especially for $c_6$), and $c_3$'s low $D^2_{\rm MD}$ value and tendency to have $\mathbf{D}_{\rm PC}$ values close to 0 in the diagnostics.

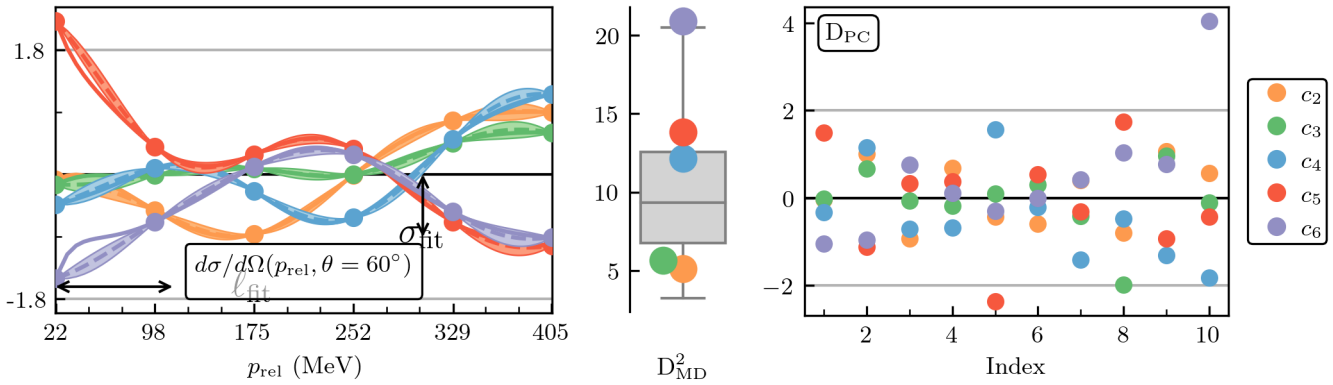



FIG. 11. Figures here are generated with the same choices as those in Fig. 10, but with $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The change in $Q$ parametrization from Fig. 10 remediates nonstationarity (except at the lowest momenta) in the coefficients, as seen in the improved diagnostics.

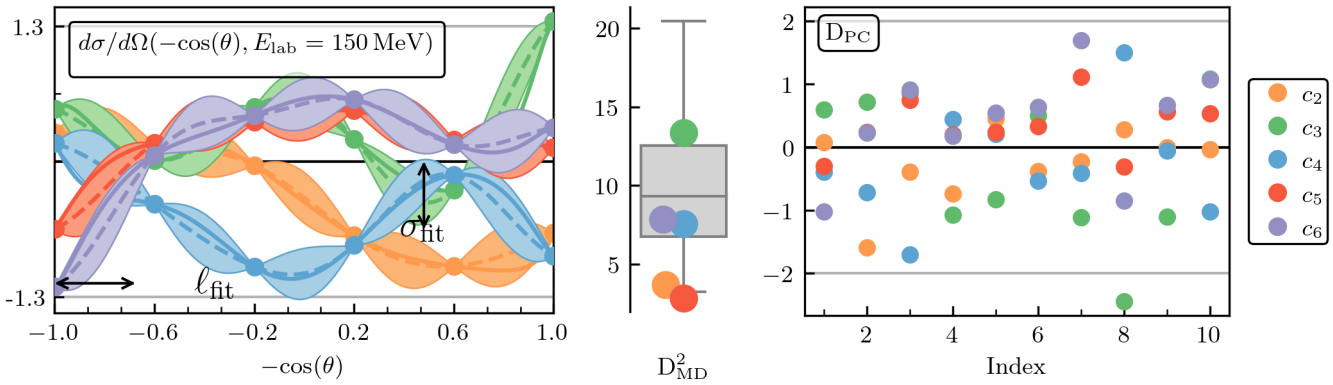



FIG. 12. Diagnostics for the differential cross section at $E_{\rm lab} = 150\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points.

consistency with our model from the coefficients and diagnostics (especially the $D^2_{MD}$ and $\mathbf{D}_{PC}$ plots) across the kinematic range and observables of interest. To show the BUQEYE model's broader applicability, additional examples of consistency are given in Appendix A. We stress that although only one potential with one regulator scheme and scale is tested here, any $\chi$EFT potential is amenable to such an analysis, which is facilitated by an accompanying Jupyter notebook.

### A. Parametrizing the expansion parameter

Our first case study examines how the choice of dimensionless expansion parameter $Q$ impacts the coefficients; we examine the differential cross section at a fixed $\theta = 60°$ in Figs. 10 and 11 as a representative example. Here and throughout this section we show the coefficients, extracted as described in the captions, plotted as solid lines in the left panel. A GP is fit at training points located at the major $x$-axis ticks (denoted as circles on each curve), with testing points located at the minor ticks. The GP mean for each curve is shown as a dashed line and the $2\sigma$ (95% confidence interval) GP bands are shown as colored regions. The width of the "bubbles" will vary depending on how the GP length scale compares to the test point spacing. In the middle and right panels of each figure are the $D^2_{MD}$ and $\mathbf{D}_{PC}$ diagnostics for this GP fit.

At first sight, coefficients in both Fig. 10 and Fig. 11 appear generally natural and the diagnostics show none of the extreme pathologies in Fig. 4, which implies that the BUQEYE model is applicable under different choices of $Q$ parametrization. However, inspection of the coefficients in Fig. 10, which uses $Q = Q_{\text{smax}}$ [see Eq. (7)], indicates nonstationarity, namely that the variance is larger and the length scale shorter at low momentum than at high momentum. This is confirmed by the $D^2_{MD}$ and $\mathbf{D}_{PC}$ plots, which show evidence of nonstationarity: The values for $c_3$ are very low ($D^2_{MD}$ plot) and cluster very close to zero ($\mathbf{D}_{PC}$ plot). Figure 11 shows improvement upon switching from parametrizing $Q$ with $Q_{\text{smax}}$ [Eq. (7)] to $Q_{\text{sum}}$ [Eq. (10)]. This is correlated with the increase in the fit variance and length scale.

As detailed in Sec. II B, we find that, once optimal values for $\Lambda_b$ and $m_{\text{eff}}$ are determined, $Q_{\text{smax}}$ and $Q_{\text{sum}}$ are generally on a par in generating sets of coefficients consistent with the BUQEYE model across many different choices of observable and other parametrizations of $p$ and $x$. However, given the compelling evidence from Sec. IV for an underlying $Q_{\text{sum}}$-like structure in the dependence of $Q$ on $p_{\text{rel}}$, we will use $Q_{\text{sum}}$ in subsequent figures.

Next we examined three parametrizations of the characteristic momentum $p$ in $Q(p, m_{\text{eff}})$: $p = p_{\text{rel}}$, $p = q_{\text{CM}}$, and $p = p_{\text{smax}}(p_{\text{rel}}, q_{\text{CM}})$. Overall, $p = p_{\text{rel}}$ performs the best; an exemplary case is shown in Fig. 12, where the apparent stationarity of the coefficients is backed up by the statistical diagnostics. In contrast, $p = q_{\text{CM}} = p_{\text{rel}}\sqrt{1-\cos(\theta)}$ is not a good choice: It has a value of roughly zero at forward angles and roughly $p_{\text{rel}}$ at backward angles, which heavily exaggerates the size of coefficients at forward angles compared to at backward angles, with predictably deleterious effects on the statistical diagnostics. Finally, $p = p_{\text{smax}}(p_{\text{rel}}, q_{\text{CM}})$ does not offer improved behavior and can even degrade the performance (see Fig. 28 in Appendix A).

### B. Parametrizing the input spaces

We also tested four different input spaces for the scattering-angle dimension $x_\theta$ of the differential cross section and spin observables. Coefficients generated with $x = -\cos(\theta)$ showed consistency with our model broadly across angles and energies. One such example is shown in Fig. 13. Because of their similar functional form, $x = -\cos(\theta)$ and $x = q^2_{\text{CM}} = p^2_{\text{rel}}(1-\cos(\theta))$ gave nearly identical results in the statistical diagnostics. We also find that $x = \theta$ provides a suitable input space in many cases, including where $x = -\cos(\theta)$ may fail to give model-consistent coefficients (see Figs. 21 and 31).

However, the choice of input space $x = q_{\text{CM}}$ can lead to problems; e.g., see Fig. 14. In this case, there is obvious nonstationarity in the length scale, which manifests as rapidly varying coefficients at high momentum and visible trumpeting in the $\mathbf{D}_{PC}$ plot.

Additionally, we tested two input spaces for the total cross section and for the lab-energy dimension of the differential cross section and spin observables: $x = E_{\text{lab}}$ and $x = p_{\text{rel}}$. We saw that the total cross section analysis in Sec. II B favored $x = p_{\text{rel}}$ as the input space. We also find cases of other observables where the stretching effect of $x = p_{\text{rel}}$ is preferred; for example, $A_y$ in Fig. 15 with $x = p_{\text{rel}}$ is more stationary than Fig. 16 with $x = E_{\text{lab}}$. While we can also find examples where $x = E_{\text{lab}}$ is preferred, our general choice is $x = p_{\text{rel}}$ because, across a wide range of hyperparameter choices, it tends to perform better.

### C. Train-test split and constraints

Though not an explicit feature of Eq. (1), the train-test split features importantly in the workflow in Table II. In some cases, the location of training and testing points may be informed by theoretical or empirical knowledge about the physical regimes where an EFT expansion fares poorly. For example, we find that the coefficients are often at least somewhat nonstationary at low energy/momentum (see Figs. 1 and 3). Even for a widely reliable choice of parametrizations and optimized values for $\Lambda_b$ and $m_{\text{eff}}$, when we train and test over all momenta in Fig. 17, we find the nonstationarity of the coefficients (mainly $c_5$) reflected starkly in the diagnostics. But when the region below 75 MeV relative momentum is omitted

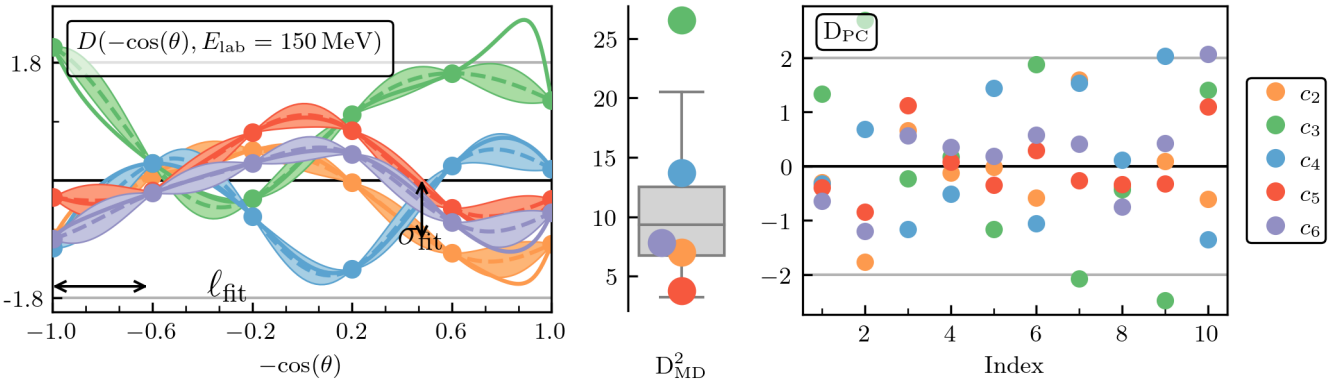



FIG. 13. Diagnostics for the spin observable $D$ at $E_{\rm lab} = 150\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points.

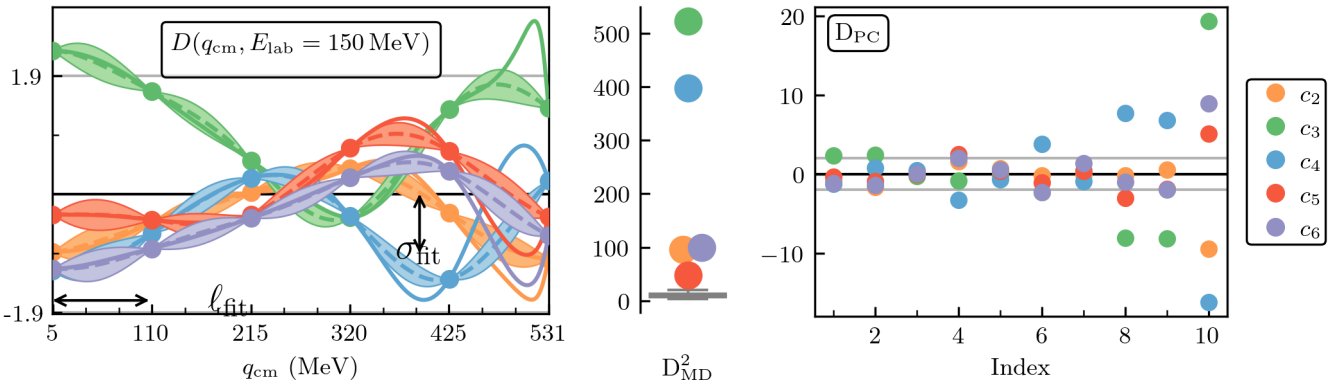



FIG. 14. Figures here are generated with the same choices as those in Fig. 13, but with $x_\theta = q_{\rm CM}$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 184\,{\rm MeV}, \Lambda_b = 650\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). Note the nonstationarity in the coefficient plots and the trumpeting in the $\mathbf{D}_{\rm PC}$ plot (especially when compared to Fig. 13), which are signs that something is amiss with the length scale.

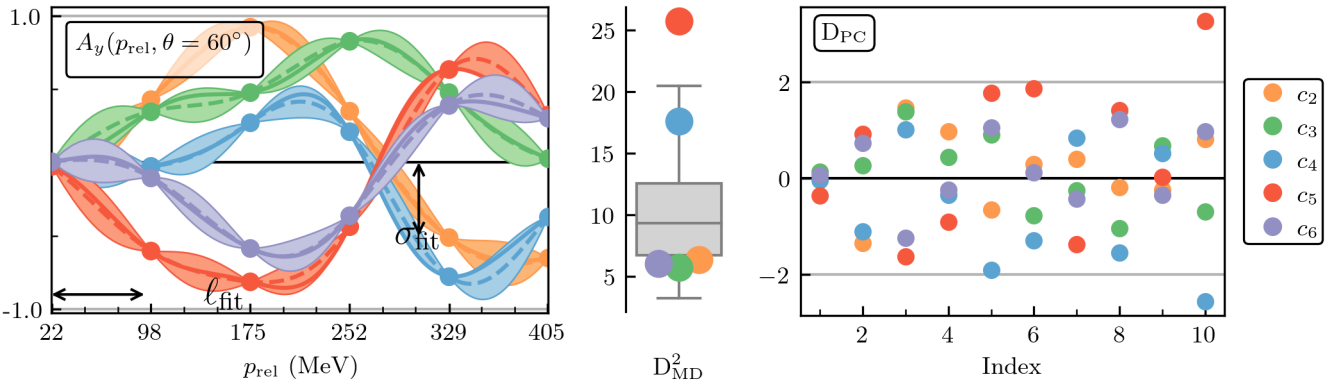



FIG. 15. Diagnostics for the spin observable $A_y$ at $\theta = 60^\circ$. Here, the coefficients are plotted with $x_E = p_{\rm rel}$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points.

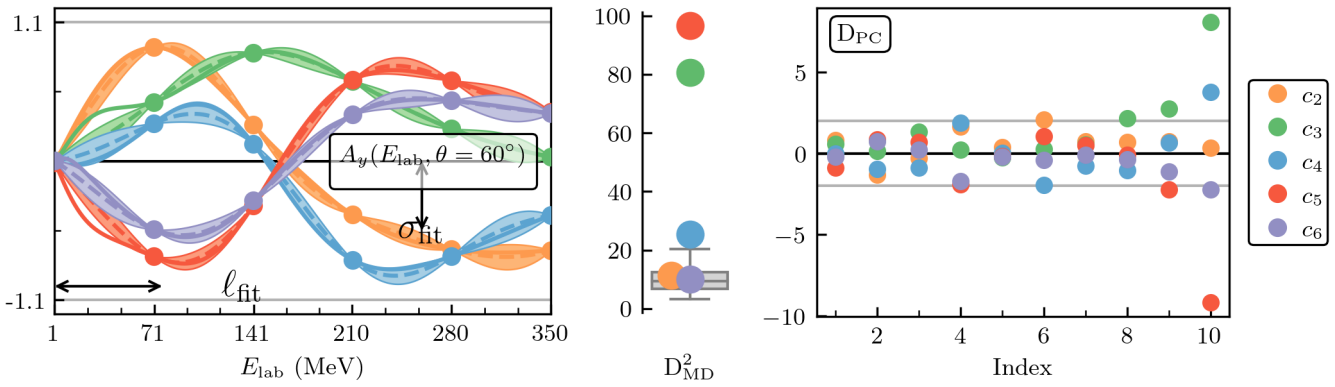



FIG. 16. Figures here are generated with the same choices as those in Fig. 15, but with $x_\theta = E_{\rm lab}$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 186\,{\rm MeV}, \Lambda_b = 610\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). Note that $c_3$ and $c_5$ seem to present particular issues in their long length scale, which the $D^2_{\rm MD}$ plot puts in stark relief.

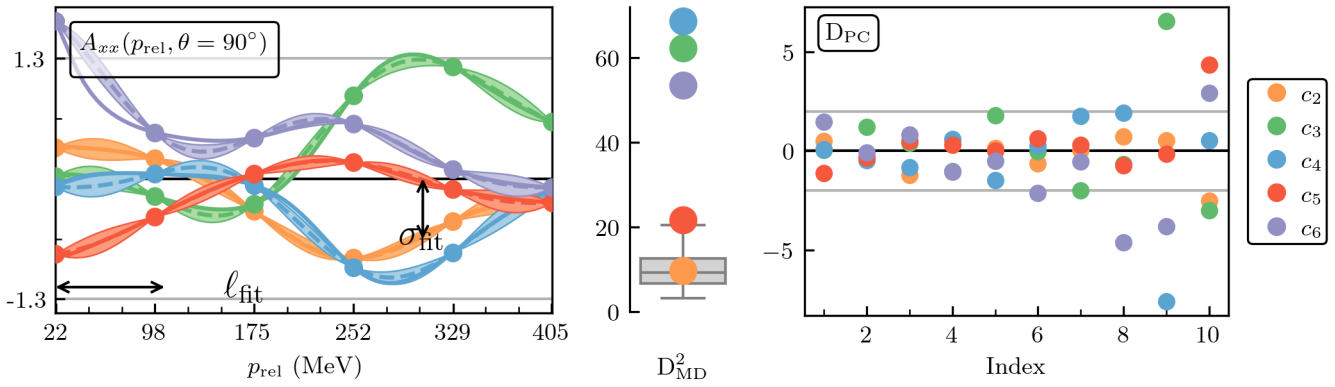



FIG. 17. Diagnostics for the spin observable $A_{xx}$ at $\theta = 90^\circ$. Here, the coefficients are plotted with $x_\theta = p_{\rm rel}$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points (starting near 22 MeV) and 10 testing points.

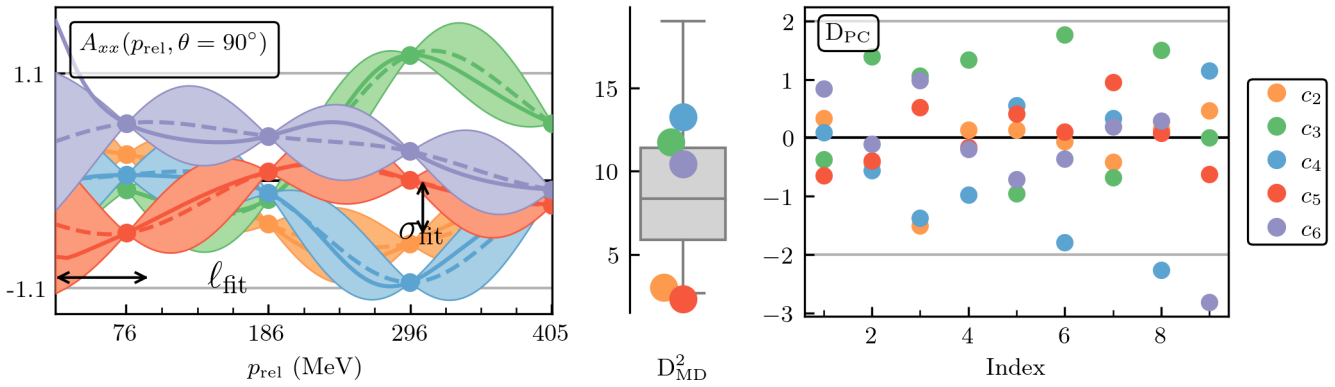



FIG. 18. Figures here are generated with the same choices as those in Fig. 17, but with 4 training points and 9 testing points. The omission of training and testing points for momenta below 75 MeV removes the trumpeting of $c_3$, $c_4$, and $c_6$ seen in Fig. 17 and so yields a consistent $D^2_{\rm MD}$ distribution.

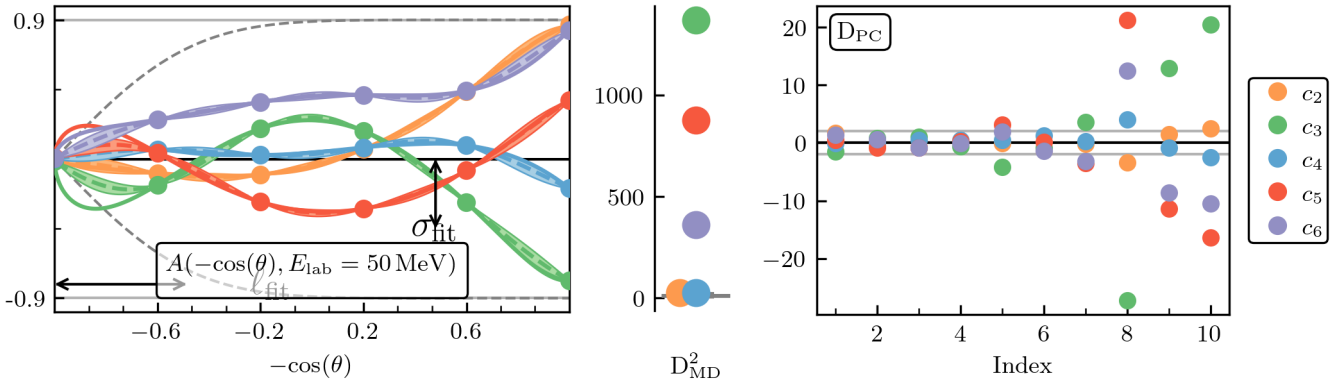



FIG. 19. Diagnostics for the spin observable $A$ at $E_{\rm lab} = 50\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. Constraints lead to bunching of coefficients at forward angle that resolves as angle increases, leading to a nonstationary length scale and diagnostics that announce nonstationarity.

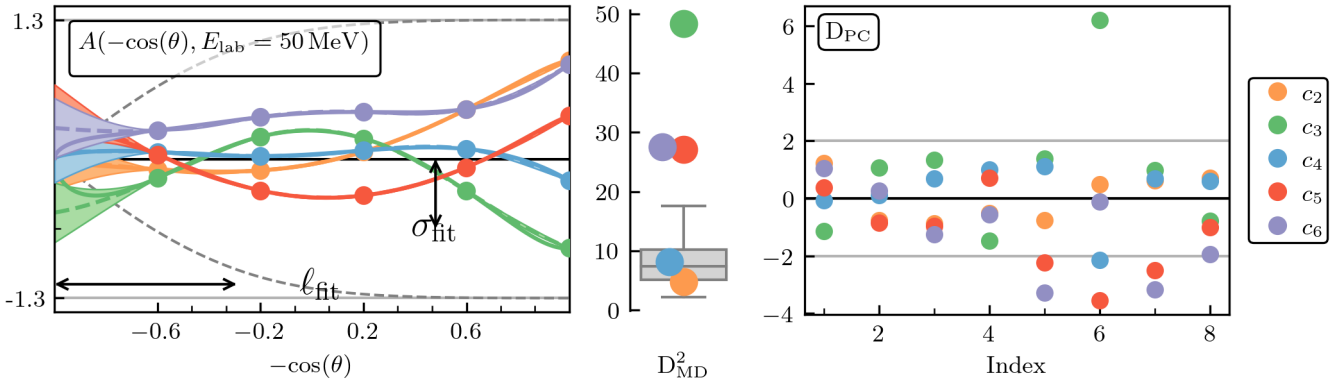



FIG. 20. Figures here are generated with the same choices as those in Fig. 19, but with 5 training points and 8 testing points. With the lack of training and testing points at forward angles, the situation of stationarity, as shown by the diagnostics, improves compared to Fig. 19 but is not ideal.

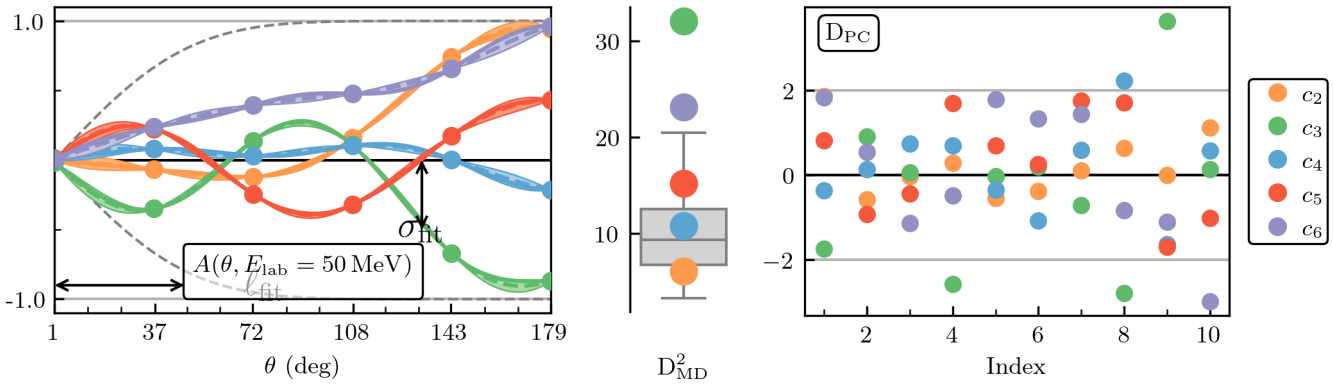



FIG. 21. Figures here are generated with the same choices as those in Fig. 19, but with $x_\theta = \theta$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 144\,{\rm MeV}, \Lambda_b = 590\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). A simple change of input space from Fig. 19 yields much improvement in the convergence pattern.

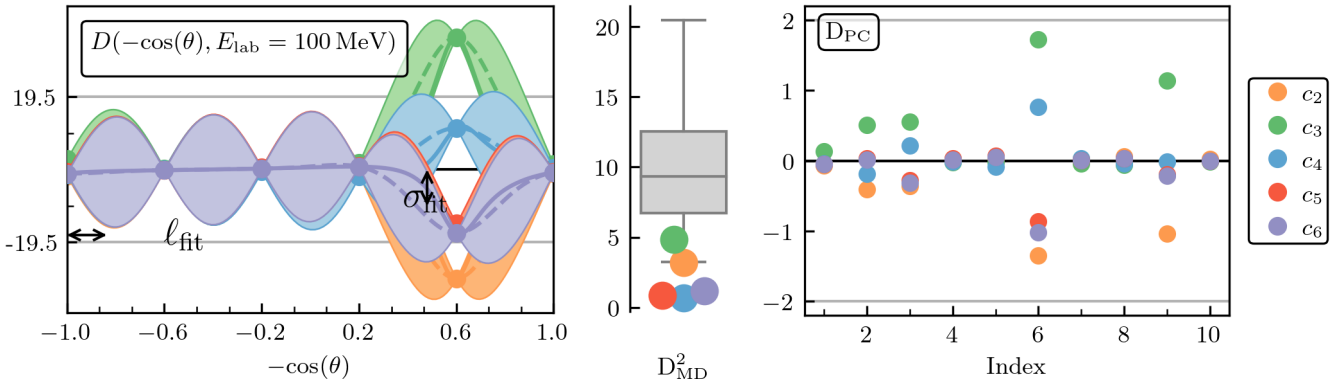



FIG. 22. Diagnostics for the spin observable $D$ at $E_{\rm lab} = 100\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. The failure arises from the fact that $y_{\rm ref}$ is set to the highest order of data that exists, which crosses zero somewhere in the input space, instead of 1. This mistake leads to extremely nonstationary and unnatural coefficients.

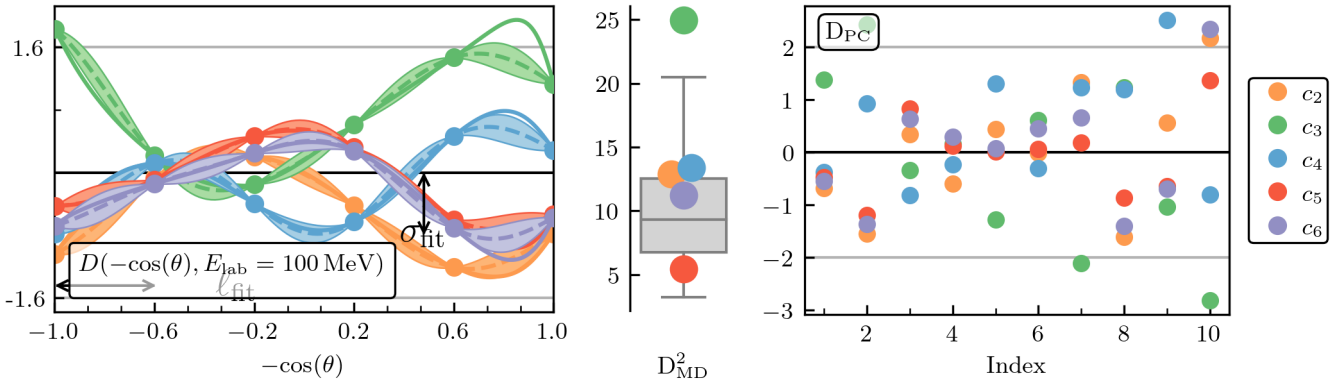



FIG. 23. Figures here are generated with the same choices as those in Fig. 22, but with $y_{\rm ref} = 1$.

from both the training and testing sets in Fig. 18, a much stronger pattern of stationarity emerges.

Aside from the exclusion of regimes where the convergence or suitability of an EFT may be in doubt, more idiosyncratic considerations may bear on the choice of train-test split. As mentioned in Sec. II A, two spin observables have constraints in their values due to time-reversal symmetry: $A(E_{\rm lab}, \theta = 0^\circ) = 0$, and $A_y(E_{\rm lab}, \theta = \{0^\circ, 180^\circ\}) = 0$. This does not affect the visual evidence of (non)stationarity in the coefficients, but it can present problems for the statistical diagnostics. Because the constraints eliminate the uncertainty that the GP can assign at those points, the diagnostics for these observables paint a very poor picture (Fig. 19) regardless of how the coefficients look. There are at least two options for remediating the situation. One might simply omit the training and testing points in the region of the constraints, as shown in Fig. 20, without changing any choices of parametrization; this approach improves consistency with our model. Alternatively, a change of input space, such as the switch from $x = \cos(\theta)$ to $x = \theta$ in Fig. 21, may also improve consistency without having to omit a portion of the input space in the train-test split. Figures 29–31 (see Appendix A) tell a similar story.

### D. Reference scale

Some care is needed in the choice of $y_{\rm ref}$ in Eq. (1). The total cross section and differential cross section are always positive and the total cross section spans many orders of magnitude in value, so it makes sense to set $y_{\rm ref}$ equal to the values of the highest calculated order, as we do in all cases in this paper. However, spin observables ($D$, $A_{xx}$, $A_{yy}$, $A$, and $A_y$) can take positive and negative values over a given input space, which makes this approach to $y_{\rm ref}$ problematic, as dividing by the set of data for the highest calculated order when it has a zero-crossing can lead to divergences. An example of the deleterious consequences is seen in Fig. 22 for the spin observable $D$, where $y_{\rm ref}$ is set to the highest order. We strongly recommend instead setting $y_{\rm ref} = 1$ for spin observables; see Fig. 23 for the corresponding plot to Fig. 22 but with $y_{\rm ref} = 1$.

## VI. PRACTICAL APPLICATIONS

Here we consider two applications of our analysis workflow.

### A. Treatment of N$^4$LO+

The SMS potential includes five complete orders (LO through N$^4$LO) as well as an additional incomplete order known as N$^4$LO+, which contains numerically important corrections to $D$- and $F$-wave scattering processes. But is N$^4$LO+ subject to the same power-counting scheme as the lower, complete orders? We can essay an answer to that question by consulting coefficient plots and their associated statistical diagnostics. Frequent, pronounced inconsistency with our model across different sets of input parameters would be taken as a sign that this order is not compliant with our power-counting paradigm (however much the inclusion of N$^4$LO+ physics makes observable predictions more accurate), but overall compliance with our model would indicate that N$^4$LO+ should be treated in the same fashion as a complete order. We examine what happens when the sixth-order (i.e., N$^4$LO+) coefficient $c_6$ is omitted from the analysis in cases of inconsistency with our model.

The vast majority of parametrization choices lead to model-consistency when $c_6$ is included. In those cases the omission of $c_6$ leaves that consistency intact and makes little difference. There are cases where model-inconsistency is observed and $c_6$ is the culpable coefficient (e.g., see Fig. 24); there, omitting it from the GP model of the coefficients can remediate the situation (e.g., see Fig. 25). However, care is required in this exercise: As the highest-order coefficient under consideration, $c_6$ is most susceptible to wrong-sizing by choices of inappropriate values for $Q$. This sensitivity should caution against definitive conclusions that $c_6$ is a pathological order *per se*. Since N$^4$LO+ does not seem significantly more likely than the other orders to be pathological and its omission does not decisively affect consistency, we conclude that it ought to be treated, at least provisionally, as a full order subject to the usual power-counting scheme.

### B. Do D$_{\rm CI}$ plots work?

In Sec. III C we introduced the D$_{\rm CI}$ (weather) plot to assess how well the variance of the GP that we fit to our training data captures the validation data. This purpose is fundamental to the BUQEYE model's goal: to obtain statistically rigorous error bands for $\chi$EFT predictions of physical observables. To do this, we need to check whether the error band from the fitted GP's variance truly encompasses the assumed percentage of the validation data; i.e., that our model is working as expected. Furthermore, it is important to verify that the D$_{\rm CI}$ plots actually concur with the other diagnostics.

Here we offer two examples from this paper and highlight the insight that D$_{\rm CI}$ plots can share. Credible interval plots can point out cases in which the EFT error bands are assessed too conservatively or not conservatively enough. An example of error bands that are too conservative occurs in the comparison of Fig. 28, in which the characteristic momentum is parametrized by $p = p_{\rm smax}(p_{\rm rel}, q_{\rm CM})$, and Fig. 12, where the more proper choice $p = p_{\rm rel}$ is made. Figure 26 shows that when the change is made, the assessment of the error goes from too conservative (which corresponds to curves in the upper-left of the weather plot) to more evenhanded (which cor-

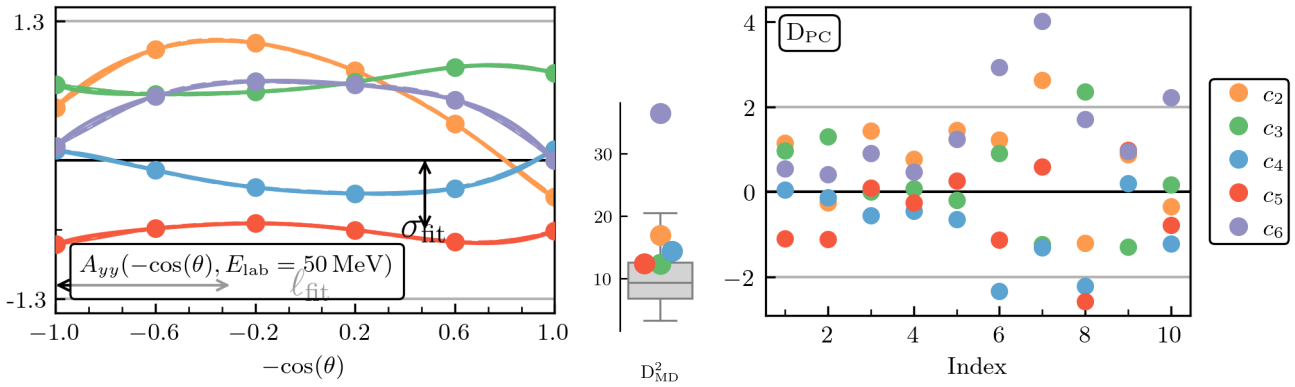



FIG. 24. Diagnostics for the spin observable $A_{yy}$ at $E_{\rm lab} = 50\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. Note that $c_6$'s $D^2_{\rm MD}$ value shows it to be an outlier.

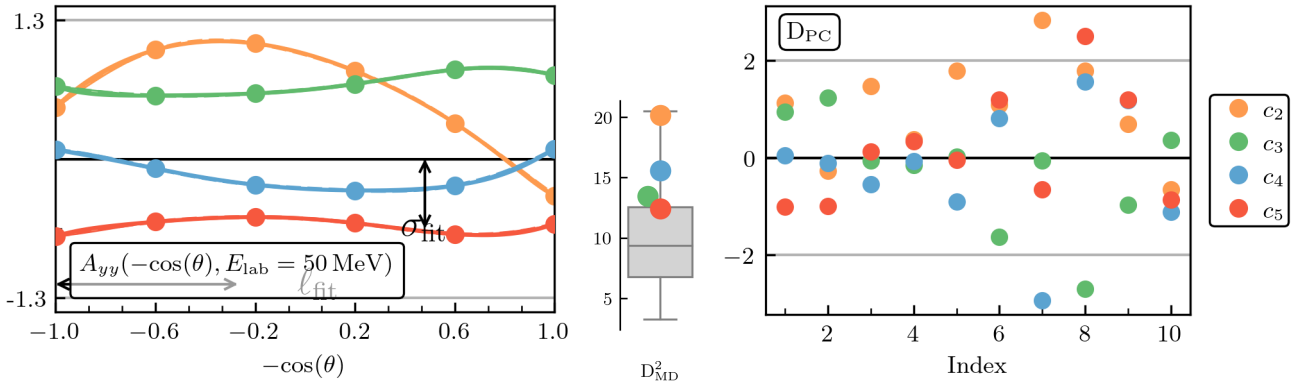



FIG. 25. Diagnostics for the spin observable $A_{yy}$ at $E_{\rm lab} = 50\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 120\,{\rm MeV}, \Lambda_b = 528\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. The coefficient function $c_6$ is omitted from all calculations. Note that this omission results in improvements in consistency shown in the diagnostics.

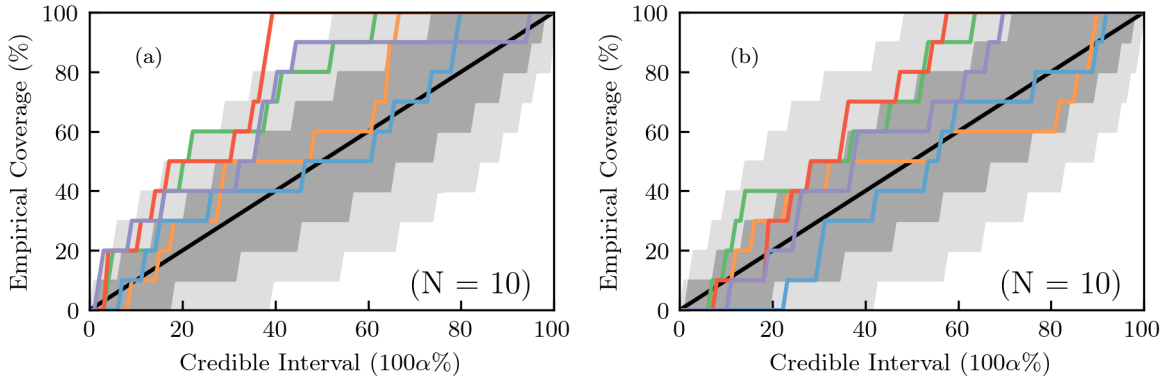



FIG. 26. Plots of the credible intervals ("weather plots") corresponding to the coefficients of the differential cross section at $E_{\rm lab} = 150\,{\rm MeV}$ in Fig. 28 (a) and Fig. 12 (b). The concentration of curves above and to the left of the black midline in the lefthand figure is an indication that the truncation error is being overestimated and the error model is too conservative. Those on the righthand side track the midline better, within the shaded confidence bounds, and thus show a better agreement between estimated and actual error bars.

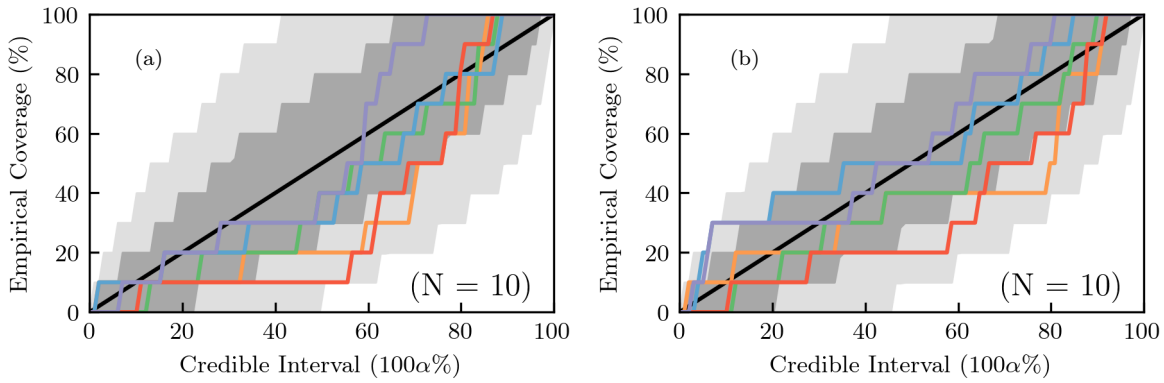



FIG. 27. Plots of the credible intervals ("weather plots") corresponding to the coefficients of the spin observable $A_y$ at $\theta = 60^\circ$ MeV in Fig. 16 (a) and Fig. 15 (b). The concentration of curves below and to the right of the black midline in the lefthand figure is an indication that the truncation error is being underestimated and the error model is not conservative enough. Those on the righthand side track the midline better, within the shaded confidence bounds, and thus show a better agreement between estimated and actual error bars.

responds to curves along the midline). For an example of error bands that are not conservative enough, see Fig. 27, which is based on Figs. 16 and 15. Here, the curves do the opposite: They move from the lower right to the midline, which signals that the uncertainty is more properly quantified with the choice of $x_E = p_{\rm rel}$ than $x_E = E_{\rm lab}$. The $\mathrm{D_{CI}}$ plots corresponding to the other graphical and statistical figures in Sec. V can be found in a Jupyter notebook with the other files associated with this paper [42].

## VII. SUMMARY AND OUTLOOK

A full Bayesian parameter estimation of the low-energy constants (LECs) that characterize interactions in chiral effective field theory ($\chi$EFT) requires rigorous assessment of theoretical errors, which in turn requires assessing whether and when a correlated truncation-error model is suitable for the data at hand. We have shown using graphical and statistical diagnostics that the semilocal momentum-space $\chi$EFT nucleon-nucleon (NN) potential of Reinert, Krebs, and Epelbaum with cutoff 500 MeV [11], with appropriate choices of Gaussian process (GP) parametrizations, shows order-by-order convergence consistent with the BUQEYE model.

This complements the recent results of Svennson *et al.*, who fitted two-dimensional GPs to the $\mathrm{N^2LO}$ coefficient $c_3$ in a $\chi$EFT treatment of NN scattering that included an explicit $\Delta(1232)$ isobar degree of freedom. Svennson *et al.* did this for 15 different $np$ scattering observables in the energy range 30 MeV $\leq T_{\rm lab} \leq$ 290 MeV and found that a stationary-GP description of $c_3$ passed statistical consistency checks in every case.

Both our findings and those of Ref. [30] imply that robust statistical estimates of the truncation error can be made. Indeed Svennson *et al.* inserted the correlated $\chi$EFT truncation error derived from their GP fit to $c_3$ in the likelihood they used to estimate the EFT LECs. We illustrated some more basic applications of a correlated EFT error model in Sec. VI.

In demonstrating the BUQEYE model's applicability, we gave guidance on where to start when choosing GP parametrizations, the signs of (in)consistency in a GP description that can be seen in statistical diagnostics, what those signs point to, and how to iterate through a workflow to assess the robustness of the parametrization choices and train-test split. We examined coefficients corresponding to several different EFT orders, and so were able to employ the requirement that they be described by a common GP to infer a $\chi$EFT breakdown scale $\Lambda_b$ in the range 500–700 MeV and an EFT soft scale $m_{\rm eff}$ in the range 100–200 MeV. The posteriors for these scales have 68% intervals narrower than this if specific GP parametrizations and forms of the EFT expansion parameter are invoked (see Table III).

These lower values of $\Lambda_b$ and $m_{\rm eff}$ contrasted with those extracted from an analysis of the expansion parameter $Q$ that included correlations in scattering angle, but employed independent GPs at a number of $p_{\rm rel}$ values between 25 and 400 MeV (see Fig. 7). The range of GP variances seen across $p_{\rm rel}$ in this pointwise-in-$p_{\rm rel}$ analysis strongly suggests that modeling the correlations with stationary GPs is too restrictive.

We have studied only one exemplary potential scheme and scale in this paper, but similar analyses are in progress for other potentials. An important part of this analysis will be to relax the assumption of stationary length scales and explore how nonstationary GPs might match the underlying structure in the NN observables. Additionally, we have left unsettled the question of how best to treat behavior in regimes corresponding to regions of input space where one model seems to fail, as in Fig. 17 in Sec. V C where our model fails with energy-dependent observables in the low-momentum regime. In our case, we assessed with statistical diagnostics how often our model was consistent with the data including and

excluding training and testing points from that region, presented both, and compared them, but that is not the only solution.

Future work will build on the insight gleaned from the present work and Ref. [30] to implement full Bayesian analyses of $\chi$EFT for nuclear observables. This will be facilitated by recent advances in devising emulators that drastically reduce calculation times for repeated Monte Carlo sampling for Bayesian methods (e.g., see [47–51]). The deficiencies we observed here at relative momenta well below $m_{\rm eff}$ motivates applying the BUQEYE model to pionless EFT [52], which is tailor-made to reproduce observables in this momentum regime. It may be feasible to use Bayesian model mixing to statistically combine chiral and pionless EFT predictions to better reproduce observables across a full range of momenta.

## ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of S. Wesolowski to an earlier version of this work. We thank Christian Forssén, Mostofa Hisham, and Simon Sundberg for useful feedback on the manuscript. The work of PJM and RJF was supported in part by the National Science Foundation Award Nos. PHY-1913069 and PHY-2209442 and the NUCLEI SciDAC Collaboration under U.S. Department of Energy MSU subcontract no. RC107839-OSU. The work of DRP was supported by the US Department of Energy under contract DE-FG02-93ER-40756 and by the Swedish Research Council via a Tage Erlander Professorship (Grant No 2022-00215). The work of RJF, DRP, and MTP was supported in part by the National Science Foundation CSSI program under Award No. OAC-2004601 (BAND Collaboration [53]).

[1] E. Epelbaum, H.-W. Hammer, and U.-G. Meißner, Rev. Mod. Phys. **81**, 1773 (2009), arXiv:0811.1338.
[2] R. Machleidt and D. R. Entem, Phys. Rept. **503**, 1 (2011), arXiv:1105.2919.
[3] H.-W. Hammer, S. König, and U. van Kolck, Rev. Mod. Phys. **92**, 025004 (2020), arXiv:1906.12122.
[4] R. J. Furnstahl, H. W. Hammer, and A. Schwenk, Few Body Syst. **62**, 72 (2021), arXiv:2107.00413 [nucl-th].
[5] I. Tews *et al.*, Few Body Syst. **63**, 67 (2022), arXiv:2202.01105 [nucl-th].
[6] E. Epelbaum, H. Krebs, and U. G. Meißner, Eur. Phys. J. A **51**, 53 (2015), arXiv:1412.0142.
[7] A. Gezerlis, I. Tews, E. Epelbaum, M. Freunek, S. Gandolfi, K. Hebeler, A. Nogga, and A. Schwenk, Phys. Rev. C **90**, 054323 (2014), arXiv:1406.0454.
[8] M. Piarulli, L. Girlanda, R. Schiavilla, R. Navarro Pérez, J. E. Amaro, and E. Ruiz Arriola, Phys. Rev. C **91**, 024003 (2015), arXiv:1412.6446.
[9] A. Ekström, G. R. Jansen, K. A. Wendt, G. Hagen, T. Papenbrock, B. D. Carlsson, C. Forssén, M. Hjorth-Jensen, P. Navrátil, and W. Nazarewicz, Phys. Rev. C **91**, 051301(R) (2015), arXiv:1502.04682.
[10] B. D. Carlsson, A. Ekström, C. Forssén, D. F. Strömberg, G. R. Jansen, O. Lilja, M. Lindby, B. A. Mattsson, and K. A. Wendt, Phys. Rev. X **6**, 011019 (2016), arXiv:1506.02466.
[11] P. Reinert, H. Krebs, and E. Epelbaum, Eur. Phys. J. A **54**, 86 (2018), arXiv:1711.08821.
[12] A. Ekström, G. Hagen, T. D. Morris, T. Papenbrock, and P. D. Schwartz, Phys. Rev. C **97**, 024332 (2018), arXiv:1707.09028.
[13] D. R. Entem, R. Machleidt, and Y. Nosyk, Phys. Rev. C **96**, 024004 (2017), arXiv:1703.05454.
[14] S. Weinberg, Phys. Lett. B **251**, 288 (1990).
[15] S. Weinberg, Nucl. Phys. B **363**, 3 (1991).
[16] R. J. Furnstahl, N. Klco, D. R. Phillips, and S. Wesolowski, Phys. Rev. C **92**, 024005 (2015), arXiv:1506.01343.
[17] J. A. Melendez, S. Wesolowski, and R. J. Furnstahl, Phys. Rev. C **96**, 024003 (2017), arXiv:1704.03308.
[18] M. Cacciari and N. Houdeau, J. High Energy Phys **09**, 039 (2011), arXiv:1105.5152.
[19] A. O'Hagan, The American Statistician **73**, 69 (2019).
[20] J. A. Melendez, R. J. Furnstahl, D. R. Phillips, M. T. Pratola, and S. Wesolowski, Phys. Rev. C **100**, 044001 (2019), arXiv:1904.10581.
[21] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, Statistical science **4**, 409 (1989).
[22] N. Cressie, Terra Nova **4**, 613 (1992).
[23] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (The MIT Press, 2006).
[24] C. Drischler, R. J. Furnstahl, J. A. Melendez, and D. R. Phillips, Phys. Rev. Lett. **125**, 202702 (2020), arXiv:2004.07232 [nucl-th].
[25] C. Drischler, J. A. Melendez, R. J. Furnstahl, and D. R. Phillips, Phys. Rev. C **102**, 054315 (2020), arXiv:2004.07805 [nucl-th].
[26] R. B. Baker, B. McClung, C. Elster, P. Maris, S. P. Weppner, M. Burrows, and G. Popa, Phys. Rev. C **106**, 064605 (2022), arXiv:2112.02442.
[27] B. Acharya and S. Bacca, (2021), arXiv:2109.13972 [nucl-th].
[28] A. Gnech, L. E. Marcucci, and M. Viviani, (2023), arXiv:2305.07568 [nucl-th].
[29] J. A. Melendez, R. J. Furnstahl, H. W. Grießhammer, J. A. McGovern, D. R. Phillips, and M. T. Pratola, Eur. Phys. J. A **57**, 81 (2021), arXiv:2004.11307 [nucl-th].
[30] I. Svensson, A. Ekström, and C. Forssén, (2023), arXiv:2304.02004 [nucl-th].
[31] S. Wesolowski, R. J. Furnstahl, J. A. Melendez,

and D. R. Phillips, J. Phys. G **46**, 045102 (2019), arXiv:1808.08211.
[32] S. Wesolowski, I. Svensson, A. Ekström, C. Forssén, R. J. Furnstahl, J. A. Melendez, and D. R. Phillips, Phys. Rev. C **104**, 064001 (2021), arXiv:2104.04441 [nucl-th].
[33] I. K. Alnamlah, E. A. Coello Pérez, and D. R. Phillips, Front. in Phys. **10**, 901954 (2022), arXiv:2203.01972 [nucl-th].
[34] M. Poudel and D. R. Phillips, J. Phys. G **49**, 045102 (2022), [Erratum: J.Phys.G 49, 099601 (2022)], arXiv:2110.01451 [nucl-th].
[35] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Adaptive computation and machine learning series (University Press Group Limited, Cambridge, MA, 2006).
[36] P. La France and P. Winternitz, Journal de Physique **41**, 1391 (1980).
[37] P. Blanchard, D. J. Higham, and N. J. Higham, IMA Journal of Numerical Analysis **41**, 2311 (2020), https://academic.oup.com/imajna/article-pdf/41/4/2311/40758053/draa038.pdf.
[38] J. Bystricky, F. Lehar, and P. Winternitz, J. Phys.(France) **39**, 1 (1978).
[39] E. Epelbaum *et al.*, Eur. Phys. J. A **56**, 92 (2020), arXiv:1907.03608 [nucl-th].
[40] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*, 3rd ed. (CRC Press, 2013).
[41] L. S. Bastos and A. O'Hagan, Technometrics **51**, 425 (2009).
[42] P. Millican, "`modern_nn_potentials`," Python package available for download from Github.
[43] R. Andrae, T. Schulze-Hartung, and P. Melchior, (2010), arXiv:1012.3754.
[44] N. Silver, *The Signal and the Noise: Why So Many Predictions Fail-but Some Don't* (Penguin Publishing Group, 2012).
[45] V. Kejzlar, L. Neufcourt, and W. Nazarewicz, Sci. Rep. **13**, 19600 (2023), arXiv:2311.01596 [stat.ME].
[46] E. Epelbaum, PoS **CD2018**, 006 (2019).
[47] C. Drischler, J. A. Melendez, R. J. Furnstahl, A. J. Garcia, and X. Zhang, Front. Phys. **10**, 92931 (2023), supplemental, interactive Python code can be found on the companion website `https://github.com/buqeye/frontiers-emulator-review`, arXiv:2212.04912.
[48] P. Giuliani, K. Godbey, E. Bonilla, F. Viens, and J. Piekarewicz, Front. Phys. **10** (2023), 10.3389/fphy.2022.1054524, arXiv:2209.13039.
[49] J. A. Melendez, C. Drischler, R. J. Furnstahl, A. J. Garcia, and X. Zhang, J. Phys. G **49**, 102001 (2022), arXiv:2203.05528 [nucl-th].
[50] D. Odell, P. Giuliani, K. Beyer, M. Catacora-Rios, M. Y. H. Chan, E. Bonilla, R. J. Furnstahl, K. Godbey, and F. M. Nunes, (2023), arXiv:2312.12426 [physics.comp-ph].
[51] T. Duguet, A. Ekström, R. J. Furnstahl, S. König, and D. Lee, (2023), arXiv:2310.19419 [nucl-th].
[52] J. Bub and et al., In preparation.
[53] Bayesian Analysis of Nuclear Dynamics (BAND) Framework project (2020) `https://bandframework.github.io/`.

### Appendix A: Additional examples

Following the discussion in Sec. V A, we include Fig. 28 to show the lack of improvement when the characteristic momentum in $Q(p, m_{\text{eff}})$ is parametrized by $p = p_{\text{smax}}(p_{\text{rel}}, q_{\text{CM}})$ instead of $p = p_{\text{rel}}$.

Additionally, we provide in Figs. 29–31 an exploration of handling constrained observables along the same lines as Figs. 19–21 (see Sec. V C).

To demonstrate further the wide applicability of the BUQEYE model, we also have included here examples of when consistency can be observed with the BUQEYE model in the cases of the spin observable $D$ sliced in energy (Fig. 32), and the spin observable $A_{xx}$ sliced in energy (Fig. 33).

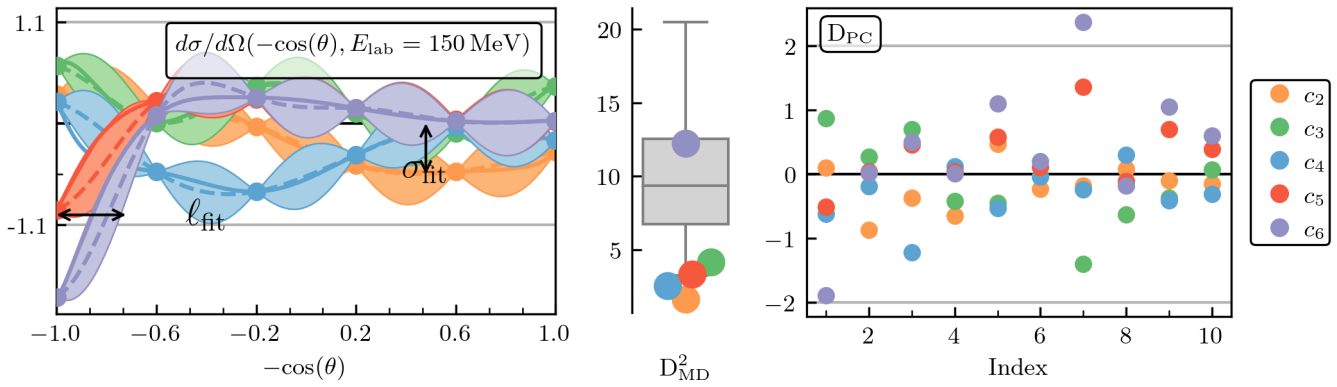



FIG. 28. Diagnostics for the differential cross section at $E_{\rm lab} = 150\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm smax}(p_{\rm rel}, q_{\rm CM}), m_{\rm eff} = 172\,{\rm MeV}, \Lambda_b = 660\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. The choice to change $p = p_{\rm rel}$, as in Fig. 28, to $p = p_{\rm smax}(p_{\rm rel}, q_{\rm CM})$ fails and even backfires by flattening the coefficients past the second training point and causing the length scale to be underestimated (see Fig. 4, "Understimated $\ell$"), with predictable results that are especially visible in the cluster of very low values for the $D^2_{\rm MD}$ plot.

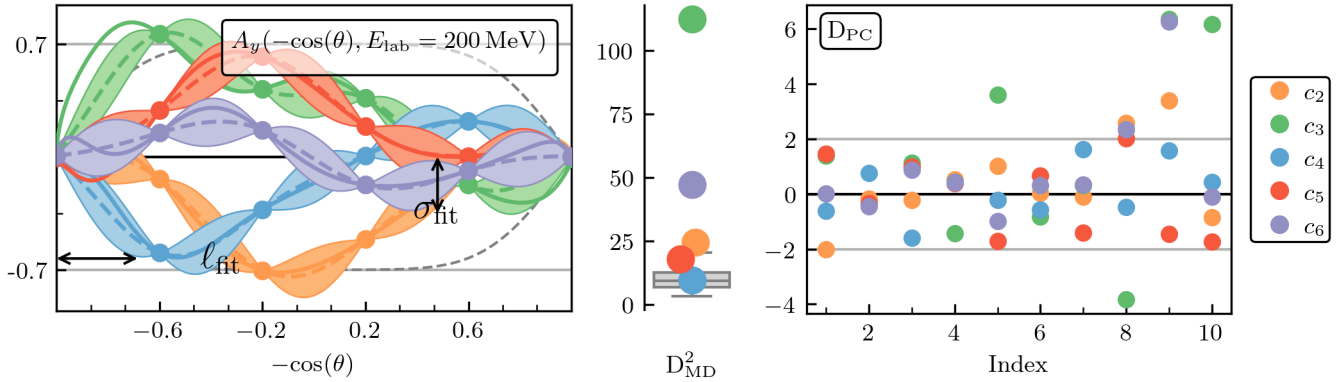



FIG. 29. Diagnostics for the spin observable $A_y$ at $E_{\rm lab} = 200\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points. Constraints lead to bunching of coefficients at forward and backward angles that resolves for angles between these extremes, leading to a nonstationary length scale and diagnostics that announce nonstationarity.

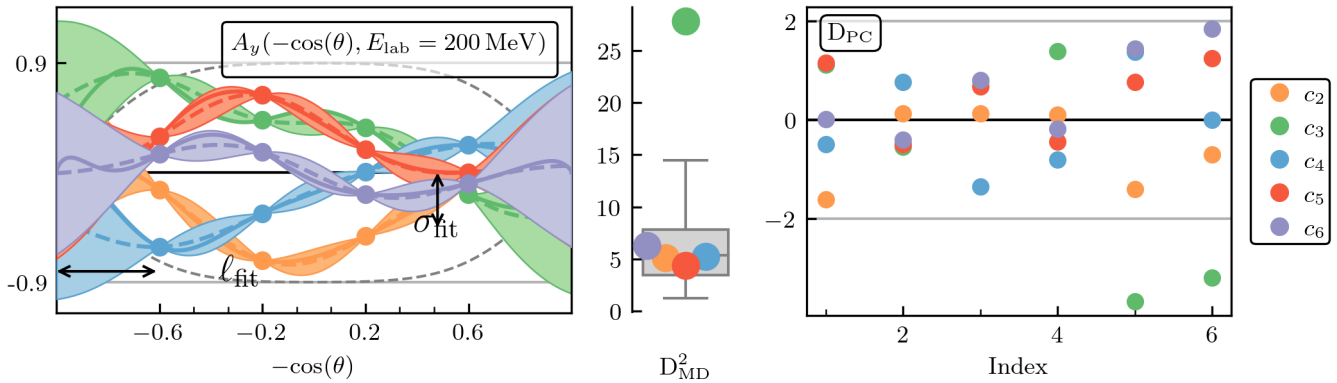



FIG. 30. Figures here are generated with the same choices as those in Fig. 29, but with 4 training points and 6 testing points. With the lack of training and testing points at forward and backward angles, the situation of stationarity, as shown by the diagnostics, improves but is not ideal.

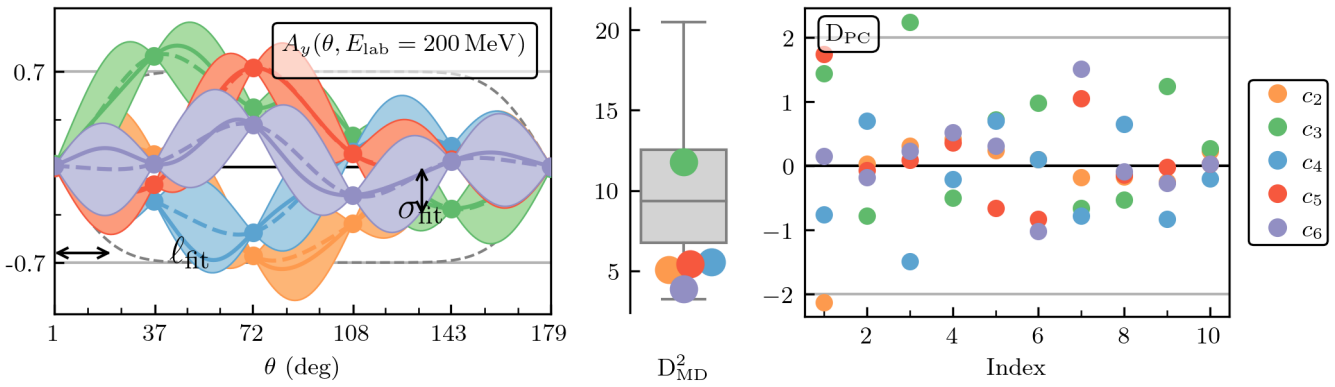



FIG. 31. Figures here are generated with the same choices as those in Fig. 29, but with $x_\theta = \theta$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 144\,{\rm MeV}, \Lambda_b = 590\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The omission of training and testing points at very forward and very backward angles is salutary for the convergence pattern, as seen in the statistical diagnostics.

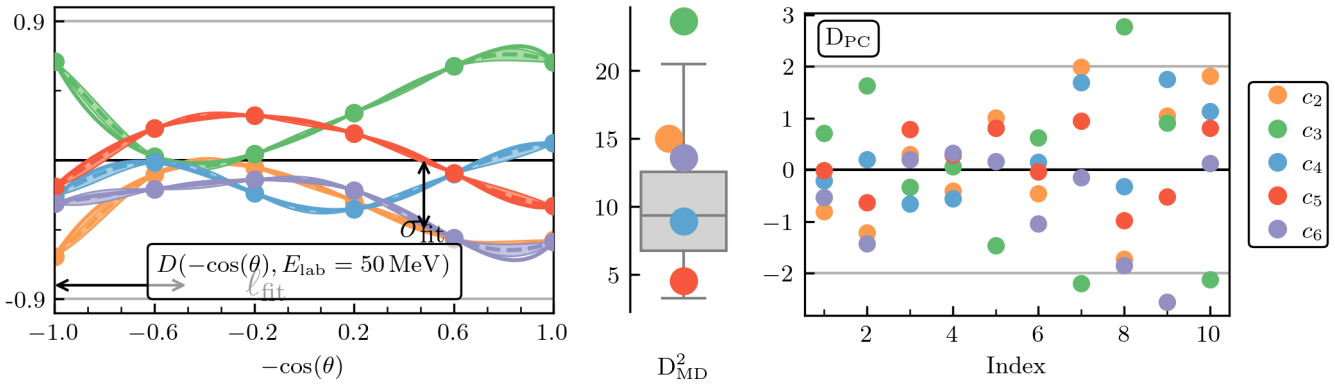



FIG. 32. Diagnostics for the spin observable $D$ at $E_{\rm lab} = 50\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points.

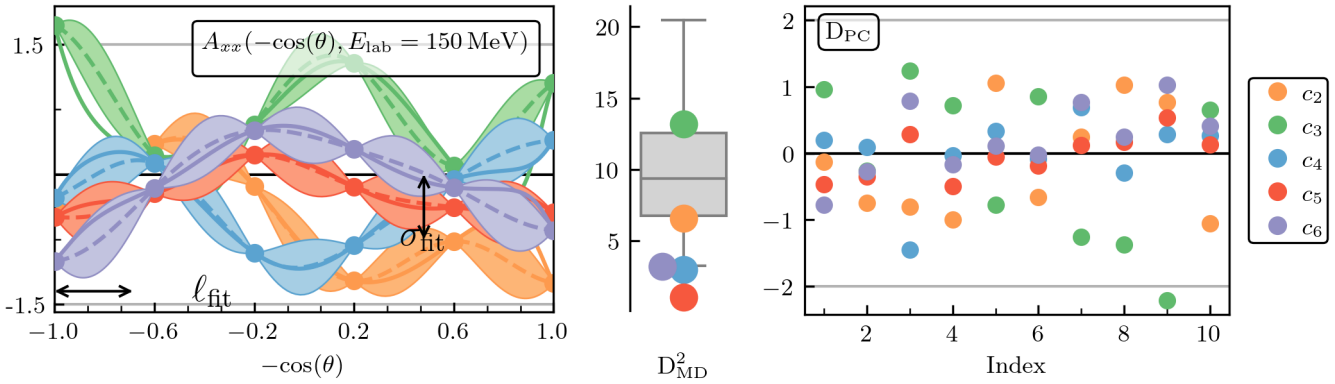



FIG. 33. Diagnostics for the spin observable $A_{xx}$ at $E_{\rm lab} = 150\,{\rm MeV}$. Here, the coefficients are plotted with $x_\theta = -\cos(\theta)$ and $Q = Q_{\rm sum}(p = p_{\rm rel}, m_{\rm eff} = 138\,{\rm MeV}, \Lambda_b = 570\,{\rm MeV})$ (optimal values of $m_{\rm eff}$ and $\Lambda_b$ from Table III). The statistical diagnostics are calculated with 6 training points and 10 testing points.