EPJ Nuclear
Sciences
& Technologies

REGULAR ARTICLE

OPEN ACCESS

Estimating model bias over the complete nuclide chart with sparse Gaussian processes at the example of INCL/ABLA and double-differential neutron spectra

Georg Schnabel[*]

Irfu, CEA, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

Abstract. Predictions of nuclear models guide the design of nuclear facilities to ensure their safe and efficient operation. Because nuclear models often do not perfectly reproduce available experimental data, decisions based on their predictions may not be optimal. Awareness about systematic deviations between models and experimental data helps to alleviate this problem. This paper shows how a sparse approximation to Gaussian processes can be used to estimate the model bias over the complete nuclide chart at the example of inclusive double-differential neutron spectra for incident protons above 100 MeV. A powerful feature of the presented approach is the ability to predict the model bias for energies, angles, and isotopes where data are missing. The number of experimental data points that can be taken into account is at least in the order of magnitude of 10^4 thanks to the sparse approximation. The approach is applied to the Liège intranuclear cascade model coupled to the evaporation code ABLA. The results suggest that sparse Gaussian process regression is a viable candidate to perform global and quantitative assessments of models. Limitations of a philosophical nature of this (and any other) approach are also discussed.

1 Introduction

Despite theoretical advances, nuclear models are in general not able to reproduce all features of trustworthy experimental data. Because experiments alone do not provide sufficient information to solve problems of nuclear engineering, models are still needed to fill the gaps.

Being in need of reliable nuclear data, a pragmatic solution to deal with imperfect models is the introduction of a bias term on top of the model. The true prediction is then given as the sum of model prediction and bias term. The form of the bias term is in principle arbitrary and a low order polynomial could be a possible choice. However, assumptions about how models deviate from reality are usually very vague, which makes it difficult to justify one parametrization over another one.

Methods of non-parametric statistics help to a certain extent to overcome this specification problem. In particular, Gaussian process (GP) regression (also known as Kriging, e.g., [1]) enjoys popularity in various fields, such as geostatistics, remote sensing and robotics, due to its conceptual simplicity and sound embedding in Bayesian statistics. Instead of providing a parameterized function to fit, one specifies a mean function and a covariance function.

The definition of these two quantities induces a prior probability distribution on a function space. Several standard specifications of parametrized covariance functions exist, e.g., [1, Chap. 4], whose parameters regulate the smoothness and the magnitude of variation of the function to be determined by GP regression.

The optimal choice of the values for these so-called hyperparameters is problem dependent and can also be automatically performed by methods such as marginal likelihood optimization and cross validation, e.g., [1, Chap. 5]. In addition, features of various covariance functions can be combined by summing and multiplying them. Both the automatic determination of hyperparameters and the combination of covariance functions will be demonstrated.

From an abstract viewpoint, GP regression is a method to learn a functional relationship based on samples of input-output associations without the need to specify a functional shape. GP regression naturally yields besides estimates also the associated uncertainties. This feature is essential for evaluating nuclear data because uncertainties of estimates are as important for the design of nuclear facilities as are the estimates themselves. Prior knowledge about the smoothness and the magnitude of the model bias can be taken into account. Furthermore, many existing nuclear data evaluation methods, e.g., [2–4], can be regarded as special cases. This suggests that the approach discussed in this paper can be combined with existing evaluation methods in a principled way.

* e-mail: georg.schnabel@nucleardata.com

The main hurdle for applying GP regression is the bad scalability in the number of data points N. The required inversion of an $N \times N$ covariance matrix leads to a computational complexity of N^3. This limits the application of standard GP regression to several thousand data points on contemporary desktop computers. Scaling up GP regression to large datasets is therefore a field of active research. Approaches usually rely on a combination of parallel computing and the introduction of sparsity in the covariance matrix, which means to replace the original covariance matrix by a low rank approximation, see e.g., [5].

In this paper, I investigate the sparse approximation introduced in [6] to estimate the model bias of inclusive double-differential neutron spectra over the complete nuclide chart for incident protons above 100 MeV. The predictions are computed by the C++ version of the Liège intranuclear cascade model (INCL) [7,8] coupled to the Fortran version of the evaporation code ABLA07 [9]. The experimental data are taken from the EXFOR database [10].

The idea of using Gaussian processes to capture deficiencies of a model exists for a long time in the literature, see e.g., [11], and has also already been studied in the context of nuclear data evaluation, e.g., [12–14]. The novelty of this contribution is the application of GP regression to a large dataset with isotopes across the nuclide chart, which is possible thanks to the sparse approximation. Furthermore, the way GP regression is applied enables predictions for isotopes without any data.

The exemplary application of sparse GP regression in this paper indicates that the inclusion of hundred thousands of data points may be feasible and isotope extrapolations yield reasonable results if some conditions are met. Therefore, sparse GP regression is a promising candidate to perform global assessments of models and to quantify their imperfections in a principled way.

The structure of this paper is as follows. Section 2 outlines the theory underlying sparse GP regression. In particular, Section 2.1 provides a succinct exposition of standard GP regression, Section 2.2 sketches how to construct a sparse approximation to a GP and how this approximation is exploited in the computation, and Section 2.3 explains the principle to adjust the hyperparameters of the covariance function based on the data.

The application to INCL/ABLA and inclusive double-differential neutron spectra is then discussed in Section 3. After a brief introduction of INCL and ABLA in Section 3.1, the specific choice of covariance function is detailed in Section 3.2. Some details about the hyperparameter adjustment are given in Section 3.3 and the results of GP regression are shown and discussed in Section 3.4.

2 Method

2.1 GP regression

GP regression, e.g., [1], can be derived in the framework of Bayesian statistics under the assumption that all probability distributions are multivariate normal. Let the vector $\boldsymbol{y}_1$ contain the values at the locations $\{\boldsymbol{x}_i^p\}$ of interest and $\boldsymbol{y}_2$ the observed values at the locations $\{\boldsymbol{x}_j^o\}$.

For instance, in the application in Section 3, the elements in the vector $\boldsymbol{y}_1$ represent the relative deviations of the "truth" from the model predictions for neutron spectra at angles and energies of interest. The vector $\boldsymbol{y}_2$ contains the relative deviations of the available experimental data from the model predictions for neutron spectra at the angles and energies of the experiments. The underlying assumption is that the experimental measurements differ from the truth by an amount compatible with their associated uncertainties. If this assumption does not hold, model bias should be rather understood as a combination of model *and* experimental bias.

Given a probabilistic relationship between the vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, i.e. we know the conditional probability density function (pdf) of $\boldsymbol{y}_2$ given $\boldsymbol{y}_1$, $\rho(\boldsymbol{y}_2|\boldsymbol{y}_1)$ (e.g., because of continuity assumptions), the application of the Bayesian update formula yields

$$\rho(\boldsymbol{y}_1|\boldsymbol{y}_2) \propto \rho(\boldsymbol{y}_2|\boldsymbol{y}_1)\rho(\boldsymbol{y}_1). \tag{1}$$

The occurring pdfs are referred to as posterior $\rho(\boldsymbol{y}_1|\boldsymbol{y}_2)$, likelihood $\rho(\boldsymbol{y}_2|\boldsymbol{y}_1)$, and prior $\rho(\boldsymbol{y}_1)$. The posterior pdf represents an improved state of knowledge.

The form of the pdfs in the Bayesian update formula can be derived from the joint distribution $\rho(\boldsymbol{y}_1,\boldsymbol{y}_2)$. In the following, we need the multivariate normal distribution

$$\mathcal{N}(\boldsymbol{y}|\mu, K) = \frac{1}{\sqrt{(2\pi)^N \det K}}$$
$$\times \exp\left(-\frac{1}{2}(\boldsymbol{y}-\mu)^T K^{-1}(\boldsymbol{y}-\mu)\right), \tag{2}$$

which is characterized by the center vector $\boldsymbol{\mu}$ and the covariance matrix K. The dimension of the occurring vectors is denoted by N. Under the assumption that all pdfs are multivariate normal and centered at zero, the joint distribution of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ can be written as

$$\rho(\boldsymbol{y}_1,\boldsymbol{y}_2) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{y}_1\\\boldsymbol{y}_2\end{pmatrix}\middle|\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}K_{11} & K_{21}^T\\K_{21} & K_{22}\end{pmatrix}\right). \tag{3}$$

The compound covariance matrix contains the blocks K_{11} and K_{22} associated with $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, respectively. The block K_{21} contains the covariances between the elements of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. Centering the multivariate normal pdf at zero is a reasonable choice for the estimation of model bias. It means that an unbiased model is a priori regarded as the most likely option.

The posterior pdf is related to the joint distribution by

$$\rho(\boldsymbol{y}_1|\boldsymbol{y}_2) = \frac{\rho(\boldsymbol{y}_1,\boldsymbol{y}_2)}{\int \rho(\boldsymbol{y}_1,\boldsymbol{y}_2)d\boldsymbol{y}_1}. \tag{4}$$

The solution for a multivariate normal pdf is another multivariate normal pdf. For equation (3), the result is given by

$$\rho(\boldsymbol{y}_1|\boldsymbol{y}_2) = \mathcal{N}(\boldsymbol{y}_1|\boldsymbol{y}'_1, K'_{11}), \tag{5}$$

with the posterior mean vector $\boldsymbol{y}'_1$ and covariance matrix K'_{11} (e.g., [1, A.3]),

$$\boldsymbol{y'}_1 = K_{21}^T K_{22}^{-1} \boldsymbol{y}_2, \qquad (6)$$

$$K'_{11} = K_{11} - K_{21}^T K_{22}^{-1} K_{21}. \qquad (7)$$

The important property of these equations is the fact that the posterior moments depend only on the observed vector $\boldsymbol{y}_2$. The introduction of new elements into $\boldsymbol{y}_1$ and associated columns and rows into K_{11} in equation (3) has no impact on the already existing values in $\boldsymbol{y'}_1$ and K'_{11}. In other words, one is not obliged to calculate all posterior expectations and covariances at once. They can be calculated sequentially.

GP regression is a method to learn a functional relationship $f(\boldsymbol{x})$ based on samples of input-output associations $\{(\boldsymbol{x}_1, f_1), (\boldsymbol{x}_2, f_2), \dots\}$. The vector $\boldsymbol{y}_2$ introduced above then contains the observed functions values of $f(\boldsymbol{x})$, i.e. $\boldsymbol{y}_2 = (f_1, f_2 \dots)^T$. Assuming all prior expectations to be zero, the missing information to evaluate equations (6) and (7) are the covariance matrices. Because predictions for $f(\boldsymbol{x})$ should be computable at all possible locations $\boldsymbol{x}$ and the same applies to observations, covariances between functions values must be available for all possible pairs $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of locations. This requirement can be met by the introduction of a so-called covariance function $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. A popular choice is the squared exponential covariance function

$$\kappa(\boldsymbol{x}_i \boldsymbol{x}_j) = \delta^2 \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^2}{2\lambda^2}\right). \qquad (8)$$

The parameter δ enables the incorporation of prior knowledge about the range functions values are expected to span. The parameter λ regulates the smoothness of the solution. The larger λ the slower the covariance function decays for increasing distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and consequently the more similar function values are at nearby locations.

If the set $\{\boldsymbol{x}_i^p\}$ contains the locations of interest and $\{\boldsymbol{x}_j^o\}$ the observed locations, the required covariance matrices in equations (6) and (7) are given by

$$K_{22} = \begin{pmatrix} \kappa(\boldsymbol{x}_1^o, \boldsymbol{x}_1^o) & \kappa(\boldsymbol{x}_1^o, \boldsymbol{x}_2^o) & \cdots \\ \kappa(\boldsymbol{x}_2^o, \boldsymbol{x}_1^o) & \kappa(\boldsymbol{x}_2^o, \boldsymbol{x}_2^o) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}, \qquad (9)$$

and

$$K_{21} = \begin{pmatrix} \kappa(\boldsymbol{x}_1^o, \boldsymbol{x}_1^p) & \kappa(\boldsymbol{x}_1^o, \boldsymbol{x}_2^p) & \cdots \\ \kappa(\boldsymbol{x}_2^o, \boldsymbol{x}_1^p) & \kappa(\boldsymbol{x}_2^o, \boldsymbol{x}_2^p) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}. \qquad (10)$$

2.2 Sparse Gaussian processes

The main hurdle for applying GP regression using many observed pairs $(\boldsymbol{x}_i, f_i)$ is the inversion of the covariance matrix k_{22} in equations (6) and (7). The time to invert this $N \times N$ matrix with N being the number of observations increases proportional to N^3. This limits the application of

standard GP regression to several thousand observations on contemporary desktop computers. Parallelization helps to a certain extent to push this limit. Another measure is the approximation of the covariance matrix by a low rank approximation. I adopted the sparse approximation described in [6], which will be briefly outlined here. For a survey of different approaches and their connections consult e.g., [5].

Suppose that we have not measured the values in $\boldsymbol{y}_2$ associated with the locations $\{\boldsymbol{x}_j^o\}$, but instead a vector $\boldsymbol{y}_3$ associated with some other locations $\{\boldsymbol{x}_k^{psi}\}$. We refer to $\boldsymbol{y}_3$ as vector of pseudo-inputs. Now we use equation (6) to determine the hypothetical posterior expectation of $\boldsymbol{y}_2$,

$$\boldsymbol{y'}_2 = \overbrace{K_{23} K_{33}^{-1}}^{S} \boldsymbol{y}_3. \qquad (11)$$

The matrices K_{23} and K_{33} are constructed analogous to equations (10) and (9), respectively, i.e. $(K_{23})_{ij} = \kappa(\boldsymbol{x}_i^o, \boldsymbol{x}_j^{psi})$ and $(K_{33})_{jk} = \kappa(\boldsymbol{x}_j^{psi}, \boldsymbol{x}_k^{psi})$ with $i = 1..N$, $j = 1..M$ and N being the number of observations and M the number of pseudo-inputs. Noteworthy, $\boldsymbol{y'}_2$ is a linear function of $\boldsymbol{y}_3$. Under the assumption of a deterministic relationship, we can replace the posterior expectation $\boldsymbol{y'}_2$ by $\boldsymbol{y}_2$. Using the sandwich formula and $\rho(\boldsymbol{y}_3) = \mathcal{N}(\boldsymbol{y}_3 | 0, K_{33})$, we get for the covariance matrix of $\boldsymbol{y}_2$

$$\tilde{K}_{22} = S K_{33} S^T = K_{23} K_{33}^{-1} K_{23}^T. \qquad (12)$$

Given that both K_{33} and K_{23} have full rank and the number of observations N is bigger than the number of pseudo-inputs M, the rank of $\tilde{K}_{22}$ equals M. The approximation is more rigid than the original covariance matrix due to the lower rank.

In order to restore the flexibility of the original covariance matrix, the diagonal matrix $D_2 = \text{diag}[K_{22} - \tilde{K}_{22}]$ is added to $\tilde{K}_{22}$. This correction is essential for the determination of the pseudo-input locations via marginal likelihood optimization as explained in [6, Sect. 3]. Furthermore, to make the approximation exhibit all properties of a GP, it is also necessary to add $D_1 = \text{diag}[K_{11} - \tilde{K}_{11}]$ to $\tilde{K}_{11} = K_{31}^T K_{33}^{-1} K_{31}$ as explained in [5, Sect. 6].

Making the replacements $K_{22} \rightarrow \tilde{K}_{22} + D_2$, $K_{21} \rightarrow \tilde{K}_{21} = K_{23} K_{33}^{-1} K_{31}$, and $K_{11} \rightarrow \tilde{K}_{11} + D_1$ in equations (6) and (7), we obtain

$$\boldsymbol{y'}_1 = \tilde{K}_{21}^T \big(\tilde{K}_{22} + D_2\big)^{-1} \boldsymbol{y}_2, \qquad (13)$$

$$K'_{11} = \tilde{K}_{11} + D_1 - \tilde{K}_{21}^T \big(\tilde{K}_{22} + D_2\big)^{-1} \tilde{K}_{21}. \qquad (14)$$

Using the Woodbury matrix identity (e.g., [1, A. 3]), these formulas can be rewritten as (e.g., [5, Sect. 6]),

$$\boldsymbol{y'}_1 = K_{31}^T \big(K_{33} + K_{23}^T D_2^{-1} K_{23}\big)^{-1} K_{23}^T D^{-1} \boldsymbol{y}_2, \qquad (15)$$

$$K'_{11} = D_1 + K_{31}^T \big(K_{33} + K_{23}^T D_2^{-1} K_{23}\big)^{-1} K_{31}. \qquad (16)$$

Noteworthy, the inversion in these expressions needs only to be performed for an $M \times M$ matrix where M is the number of pseudo-inputs. Typically, M is chosen to be in the order of magnitude of hundred. The computational cost for inverting the diagonal matrix D is negligible.

Using equations (15) and (16), the computational complexity scales linearly with the number of observations N due to the multiplication by K_{23}. This feature enables to increase the number of observations by one or two orders of magnitude compared to equations (6) and (7) in typical scenarios.

2.3 Marginal likelihood maximization

Covariance functions depend on parameters (called hyperparamters) whose values must be specified. In a full Bayesian treatment, the choice of values should not be informed by the same observations that enter into the GP regression afterwards. In practice, this ideal is often difficult to achieve due to the scarcity of available data and therefore frequently abandoned. A full Bayesian treatment may also become computationally intractable if there are too much data.

Two popular approaches to determine the hyperparameters based on the data are marginal likelihood maximization and cross validation, see e.g., [1, Chap. 5]. A general statement which approach performs better cannot be made. I decided to use marginal likelihood optimization because it can probably be easier interfaced with existing nuclear data evaluation methods.

Given a covariance function depending on some hyperparameters, e.g., $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j | \delta, \lambda)$ with hyperparameters δ and λ as in equation (8), the idea of marginal likelihood maximization is to select values for the hyperparameters that maximize the probability density for the observation vector $\rho(\boldsymbol{y}_2)$. In the case of the multivariate normal pdf in equation (3), it is given by (e.g., [1, Sect. 5.4.1])

$$\ln\rho(\boldsymbol{y}_2) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln\det K_{22} - \frac{1}{2}\boldsymbol{y}_2^T K_{22}^{-1}\boldsymbol{y}_2. \quad (17)$$

The first term is a constant, the second term is up to a constant the information entropy of the multivariate normal distribution, and the third term is the generalized χ^2-value. The maximization of this expression amounts to balancing two objectives: minimizing the information entropy and maximizing the χ^2-value.

The partial derivative of equation (17) with respect to a hyperparameter is given by (e.g., [1, Sect. 5.4.1])

$$\frac{\partial\ln\rho(\boldsymbol{y}_2)}{\partial\lambda} = -\frac{1}{2}\mathrm{Tr}\left(K_{22}^{-1}\frac{\partial K_{22}}{\partial\lambda}\right) + \frac{1}{2}\boldsymbol{y}_2^T K_{22}^{-1}\frac{\partial K_{22}}{\partial\lambda}K_{22}^{-1}\boldsymbol{y}_2, \quad (18)$$

which enables the usage of gradient-based optimization algorithms. In this paper, I use the L-BFGS-B algorithm [15] because it can deal with a large number of parameters and allows to impose restrictions on their ranges.

Due to the appearance of the determinant and the inverse of K_{22}, the optimization is limited to several thousand observations on contemporary desktop computers. However, replacing K_{22} by the approximation $\tilde{K}_{22} + D_2$ in equations (17) and (18) enables to scale up the number of observations by one or two orders of magnitude. The structure of the approximation is exploited by making use of the matrix determinant lemma, the Woodbury identity, and the trace being invariant under cyclic permutations.

The approximation to K_{22} is not only determined by the hyperparameters but also by the location of the pseudo-inputs $\{\boldsymbol{x}_k^{psi}\}$. Hyperparameters and pseudo-input locations can be jointly adjusted by marginal likelihood maximization. The number of pseudo-inputs is usually significantly larger (e.g., hundreds) than the number of hyperparameters (e.g., dozens). In addition, the pseudo-inputs are points in a potentially multi-dimensional space and their specification requires a coordinate value for each axis. For instance, in Section 3 sparse GP regression is performed in a five dimensional space with three hundred pseudo-inputs, which gives 1500 associated parameters. Because equation (18) has to be evaluated for each parameter in each iteration of the optimization algorithm, its efficient computation is important.

The mathematical details are technical and tedious and hence only the key ingredient for efficient computation will be discussed. Let x_{kl} be the lth coordinate of the kth pseudo-input. The crucial observation is that $\partial K_{33}/\partial x_{kl}$ yields a matrix in which only the lth and kth column and row contain non-zero elements. A similar statement holds for $\partial K_{23}/\partial x_{kl}$. This feature can be exploited in the multiplications and the trace computation in equation (18) (where K_{22} is substituted by $\tilde{K}_{22} + D_2$) to achieve $\mathcal{O}(NM)$ per coordinate of a pseudo-input with M being the number of pseudo-inputs, and N being the number of observations. This is much more efficient than $\mathcal{O}(dNM^2)$ for the partial derivative with respect to a hyperparameter.

3 Application

3.1 Scenario

The model bias was determined for the C++ version of the Liège INCL [7], a Monte Carlo code, coupled to the evaporation code ABLA07 [9] because this model combination performs very well according to an IAEA benchmark of spallation models [16,17] and is used in transport codes such as MCNPX and GEANT4. This suggests that the dissemination of a more quantitative performance assessment of INCL coupled to ABLA potentially helps many people to make better informed decisions. Because some model ingredients in INCL are based on views of classical physics (as opposed to quantum physics), the model is mainly used for high-energy reactions above 100 MeV.

The ability of a model to accurately predict the production of neutrons and their kinematic properties may be regarded as one of the most essential features for nuclear engineering applications. Especially for the development of the innovative research reactor MYRRHA [18] driven by a proton accelerator, these quantities need to be well predicted for incident protons.

Table 1. Summary of the double-differential $(p,X)n$ data above 100 MeV incident energy found in EXFOR [10]. No mass number is appended to the isotope name in the case of natural composition. Columns are: incident proton energy (En), range of emitted neutron energy (E_{min}, E_{max}), range of emission angles (θ_{min}, θ_{max}), and number of data points (NumPts). Energy related columns are in MeV and angles in degree. The total number of data points is 9287.

Isotope	En	E_{min}	E_{max}	θ_{min}	θ_{max}	NumPts
C	800	1.2	700	15	150	189
C	1500	1.2	1250	15	150	245
C	3000	1.2	2500	15	150	128
Na23	800	3.5	266	30	150	84
Al27	800	1.2	700	15	150	119
Al27	1000	2.5	280	15	150	223
Al27	1200	2.0	1189	10	160	404
Al27	1500	1.2	1250	15	150	129
Al27	1600	2.5	280	15	150	226
Al27	3000	1.2	2500	15	150	132
Fe	800	1.2	771	10	160	505
Fe	1200	2.0	1171	10	160	417
Fe	1500	1.2	1250	15	150	129
Fe	1600	2.0	1572	10	160	460
Fe	3000	1.2	2500	15	150	133
Cu	1000	2.5	280	15	150	227
Cu	1600	2.5	280	15	150	231
Zr	1000	2.5	280	15	150	229
Zr	1200	2.0	1189	10	160	423
Zr	1600	2.5	280	15	150	229
In	800	1.2	700	15	150	116
In	1500	1.2	1250	15	150	128
In	3000	1.2	2500	15	150	133
W	800	3.1	333	30	150	110
W	1000	2.5	280	15	150	231
W	1200	2.0	1189	10	160	413
W	1600	2.5	280	15	150	231
Pb	318	5.4	356	7	7	53
Pb	800	1.2	771	10	160	624
Pb	1000	2.5	280	15	150	231
Pb	1200	2.0	1189	10	160	563
Pb	1500	1.2	1250	15	150	249
Pb	1600	2.0	1591	10	160	691
Pb	3000	1.2	2500	15	150	131
Pb208	2000	0.4	402	30	150	170
Th232	1200	2.0	1189	10	160	351

For these reasons, I applied the approach to determine the model bias in the prediction of inclusive double-differential neutron spectra for incident protons and included almost all nuclei for which I found data in the EXFOR database [10]. Roughly ten thousand data points were taken into account. Table 1 gives an overview of the data.

3.2 Design of the covariance function

The covariance function presented in equation (8) is probably too restrictive to be directly used on double-differential spectra. It incorporates the assumption that the model bias spans about the same range for low and high emission energies. Because the neutron spectrum quickly

declines by orders of magnitude with increasing emission energy, it is reasonable to use a covariance function with more flexibility to disentangle the systematics of the model bias associated with these two energy domains.

Assuming the two covariance functions $\kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j)$, a more flexible covariance function can be constructed in the following ways (e.g., [1, Sect. 4.2.4])

$$\kappa_{1+2}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + \kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j), \qquad (19)$$

$$\kappa_{1 \times 2}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j)\kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j). \qquad (20)$$

Taking into account that the purpose of a covariance function is to compute elements of a covariance matrix, the construction in equation (19) is analogous to the possible construction of an experimental covariance matrix: An experimental covariance matrix can be assembled by adding a diagonal covariance matrix reflecting statistical uncertainties to another one with systematic error contributions. Please note that sparse GP regression as presented in this paper works only with a diagonal experimental covariance matrix. Equation (20) will be used to achieve a transition between the low and high energy domains.

To state the full covariance function used in this paper, the following covariance function on a one-dimensional input space is introduced,

$$\kappa(x, x'|\lambda) = \exp\left(-\frac{(x - x')^2}{2\lambda^2}\right). \qquad (21)$$

The meaning of the hyperparameter λ was explained below equation (8).

The inclusive double-differential neutron spectra for incident protons over the complete nuclide chart can be thought of as a function of spectrum values associated with points in a five dimensional input space. The coordinate axes are incident energy (En), mass number (A), charge number (Z), emission angle (θ) and emission energy (E). Using equation (21) we define two covariance functions κ_L and κ_H associated with the low and high energy domain of the emitted neutrons. Given two input vectors

$$\boldsymbol{x}_i = (En_i, A_i, Z_i, \theta_i, E_i)^T, \qquad (22)$$

$$\boldsymbol{x}_j = (En_j, A_j, Z_j, \theta_j, E_j)^T, \qquad (23)$$

the form of the covariance function for both κ_L and κ_H is assumed to be

$$\kappa_C(\boldsymbol{x}_i, \boldsymbol{x}_j|\delta_C, \lambda_C) = \delta_C^2 \kappa(En_i, En_j|\lambda_{En}^C)\kappa(A_i, A_j|\lambda_A^C)$$
$$\kappa(Z_i, Z_j|\lambda_Z^C)\kappa(\theta_i, \theta_j|\lambda_\theta^C)\kappa(E_i, E_j|\lambda_E^C). \qquad (24)$$

The hyperparameters λ_x^C are for the coordinate axis indicated by $x \in \{En, A, ...\}$ and can take different values for κ_L and κ_H ($C \in \{L, H\}$).

The transition between the two energy domains, which means to switch from κ_L to κ_H, is established by the logistic function

$$\sigma(\boldsymbol{x}|\boldsymbol{k}, \boldsymbol{x}_0) = \frac{1}{1 + \exp(-\boldsymbol{k}(\boldsymbol{x} - \boldsymbol{x}_0))}. \qquad (25)$$

Noteworthy, all the variables are vectors. The equation $\boldsymbol{k}(\boldsymbol{x} - \boldsymbol{x}_0) = 0$ defines a (hyper)plane with normal vector $\boldsymbol{k}$ and distance $|\boldsymbol{k}\boldsymbol{x}_0|/|\boldsymbol{k}|$ to the origin of the coordinate system. The function in equation (25) attains values close to zero for $\boldsymbol{x}$ far away from the plane on one side and close to one if far away on the other side. Within which distance to the plane the transition from zero to one occurs depends on the length of $\boldsymbol{k}$. The larger $|\boldsymbol{k}|$, the faster the transition and the narrower the window of transition around the plane.

With the abbreviations

$$\tau_L(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{k}, \boldsymbol{x}_0) = \sigma(\boldsymbol{x}_i|\boldsymbol{k}, \boldsymbol{x}_0)\sigma(\boldsymbol{x}_j|\boldsymbol{k}, \boldsymbol{x}_0), \qquad (26)$$

$$\tau_H(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{k}, \boldsymbol{x}_0) = (1 - \sigma(\boldsymbol{x}_i|\boldsymbol{k}, \boldsymbol{x}_0)) \qquad (27)$$

$$\times (1 - \sigma(\boldsymbol{x}_j|\boldsymbol{k}, \boldsymbol{x}_0)), \qquad (28)$$

the full covariance function κ_{full} is given by

$$\kappa_{\text{full}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tau_L(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{k}, \boldsymbol{x}_0)\kappa_L(\boldsymbol{x}_i, \boldsymbol{x}_j|\delta_L, \lambda_H)$$
$$+ \tau_H(\boldsymbol{x}_i, \boldsymbol{x}_j|\boldsymbol{k}, \boldsymbol{x}_0)\kappa_H(\boldsymbol{x}_i, \boldsymbol{x}_j|\delta_H, \lambda_H). \qquad (29)$$

Finally, the GP regression is not performed on the absolute difference Δ between a model prediction σ_{mod} and experimental data point σ_{exp}, but on the transformed quantity

$$\tilde{\Delta} = \frac{(\sigma_{\text{exp}} - \sigma_{mod})}{\max(\sigma_{mod}, 0.1)}. \qquad (30)$$

In words, relative differences are taken for model predictions larger than 0.1 and absolute differences scaled up by a factor of ten for model predictions below 0.1. Relative differences fluctuate usually wildly for spectrum values close to zero–especially for a Monte Carlo code such as INCL–and the switch to absolute values helps GP regression to find more meaningful solutions with a better ability to extrapolate.

Due to the number of roughly ten thousand data points, the covariance matrices computed with equation (29) were replaced by the sparse approximation outlined in Section 2.2. I introduced three hundred pseudo-inputs and placed them randomly at the locations of the experimental data. Their locations were then jointly optimized with the hyperparameters, which will be discussed in Section 3.3.

The diagonal matrix D_2 occurring in the approximation was changed to

$$\tilde{D}_2 = D_2 + B + P, \qquad (31)$$

to accommodate statistical uncertainties of the model prediction (due to INCL being a Monte Carlo code) and the experimental data. Both B and P are diagonal matrices. The matrix B contains variances corresponding to 10% statistical uncertainty for all experimental data points. The matrix P contains the estimated variances of the model predictions.

Table 2. Setup of the L-BFGS-B optimization and evolution of hyperparameters. The columns LB and UB give the lower bound and upper bound, respectively, of the parameter ranges. Columns It_i indicate the parameter values after $i \times 1000$ iterations. The column It_0 contains initial values and It_f the final values after 3500 iterations.

Name	LB	UB	It_0	It_1	It_2	It_3	It_f
δ_L	0.01	0.50	0.05	0.50	0.50	0.50	0.50
λ_{En}^L	50	1000	100	100	99	99	99
λ_A^L	10	300	100	100	101	103	103
λ_Z^L	10	200	40	39	39	40	41
λ_θ^L	10	179	50	51	57	66	68
λ_E^L	2	1000	20.0	7.3	5.0	4.7	4.8
δ_H	0.01	0.50	0.40	0.33	0.34	0.33	0.33
λ_{En}^H	50	1000	300	292	285	275	272
λ_A^H	10	300	100	107	110	114	115
λ_Z^H	10	200	40	49	50	50	49
λ_θ^H	10	179	50	62	63	64	64
λ_E^H	10	1000	20	40	41	42	43
k_c	0.1	20	0.2	0.4	0.4	0.3	0.3
x_c	2.0	500	6.0	2.6	2.3	2.4	2.8
γ	1.0	2.1	1.6	1.6	1.6	1.6	1.6

A diagonal matrix for the experimental covariance matrix B can certainly be challenged because the important systematic errors of experiments reflected in off-diagonal elements are neglected. This is at the moment a limitation of the approach.

3.3 Marginal likelihood maximization

The hyperparameters appearing in equation (29) and the locations of the three hundred pseudo-inputs $\{\boldsymbol{x}_k^{psi}\}$ determining the approximation in equation (11) were adjusted via marginal likelihood maximization described in Section 2.3. To be explicit, the hyperparameters considered were δ_L, δ_H,

$$\boldsymbol{\lambda}_L = (\lambda_{\mathrm{En}}^L, \lambda_A^L, \lambda_Z^L, \lambda_\theta^L, \delta_E^L), \tag{32}$$

$$\boldsymbol{\lambda}_H = (\lambda_{\mathrm{En}}^H, \lambda_A^H, \lambda_Z^H, \lambda_\theta^H, \delta_E^H), \tag{33}$$

and also $\boldsymbol{x}_0$ and $\boldsymbol{k}$ of the logistic function. The vector $\boldsymbol{k}$ was forced to remain parallel to the axes associated with En, A, and Z, i.e. $\boldsymbol{k} = (0, 0, 0, k_\theta, k_E)^T$. Further, polar coordinates $k_\theta = k_c \sin\gamma$, $k_E = k_c \cos\gamma$ were introduced. The direction of the vector $\boldsymbol{x}_0$ was taken equal to that of $\boldsymbol{k}$, which removes ambiguity in the plane specification without shrinking the set of possible solutions. Because of these measures, it sufficed to consider the length $x_c = |\boldsymbol{x}_0|$ as hyperparameter. Counting both hyperparameters and pseudo-inputs, 1515 parameters were taken into account in the optimization.

I employed the L-BFGS-B algorithm [15] as implemented in the *optim* function of R [19], which makes use of an analytic gradient, can deal with a large number of variables and permits the specification of range restrictions. The optimization was performed on a cluster using 25 cores and was stopped after 3500 iterations, which took about 10 h.

The obtained solution corresponds to $\chi^2/N = 1.03$ and is with a two-sided p-value of 0.04 reasonably consistent in a statistical sense. Restrictions of parameter ranges were established to introduce prior knowledge and to guide the optimization procedure. Noteworthy, lower limits on length-scales, such as λ_θ^L and λ_E^L; were introduced to counteract their dramatic reduction due to inconsistent experimental data in the same energy/angle range. Table 2 summarizes the optimization procedure. The evolution of the pseudo-inputs projected onto the (A, Z)-plane is visualized in Figure 1.

A thorough study of the optimization process exceeds the scope of this paper and hence I content myself with a few remarks. The length scales associated with the emission angle, i.e. λ_θ^C, experienced significant changes. Their increase means that the model bias is similar for emission angles far away from each other and the GP process is able to capture this similarity. Concerning the length scales associated with the emission energy, the small value $\lambda_E^L = 4.8$ compared to the larger value $\lambda_E^H = 43$ indicates that the features of the model bias of the low and high energy domain are indeed different. The most striking feature, however, is that the large length scales λ_A^C and λ_Z^C "survived" the optimization, which means that the model bias behaves similar over large regions of the nuclide charts. Examples of isotope extrapolations will be given in Section 3.4.

As a final remark, a more rigorous study of the optimization procedure is certainly necessary and there is room for improvement. This is left as future work.

3.4 Results and discussion

The values of the hyperparameters and pseudo-inputs obtained by marginal likelihood maximization were used in the covariance function in equation (29). This covariance
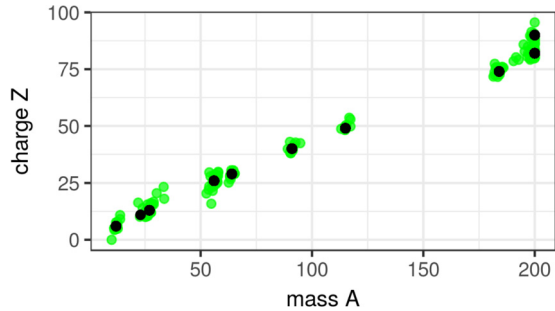
Fig. 1. Dispersion of the pseudo-inputs during the optimization procedure shown in the (A,Z)-projection. The black points are the initial positions and the green points the final positions of the pseudo-inputs.

function was employed to compute the required covariance matrices in equations (15) and (16) based on the experimental data summarized in Table 1. Because the hyperparameters are determined during hyperparameter optimization before being used in the GP regression, they will be referred to as prior knowledge.

Equations (15) and (16) enable the prediction of a plethora of spectrum values and their uncertainties for combinations of incident energy, mass number, charge number, emission angle, and emission energy. The few selected examples of predictions in Figure 2 serve as the basis to discuss general features of GP regression, the underlying assumptions, and its accuracy and validity.

How well we can interpolate between data and how far we can extrapolate beyond the data depends on the suitability of the covariance function for the problem at hand. The building block for the full covariance function in equation (29) is the one-dimensional covariance function in equation (21). Using the latter imposes the assumption that possible solutions have derivatives of any order and hence are very smooth [1, Chap.4]. Interpolations between data points included in the regression are determined by this smoothness property and values of the length scales λ_x^C.

The length scales reflect the prior assumption about similarity between spectrum values of points a certain distance away from each other. This prior assumption directly impacts the uncertainty of the predictions. The farther away a prediction is from an observation point, the higher the associated uncertainty. If a prediction is already multiples of any length scale away from all observations, the uncertainty reverts to its prior value given by either δ_L or δ_H depending on the energy domain.

In the case of the sparse approximation, the uncertainty is related to the distance to the pseudo-inputs. Because only few pseudo-inputs are located at very high and very low emission energies, the 2σ uncertainty bands in Figure 2 in those energy domains are rather large despite the presence of experimental data.

The important finding in this specific application is that the length scales related to emission angle, λ_E^C, mass number, λ_A^C, and charge number, λ_E^C, are very large. GP regression is therefore able to interpolate and extrapolate over large ranges of these axes.

The interpolation of spectrum values between angles and emission energies of an isotope with available data may be considered rather standard. For instance, one can do a χ^2-fit of a low order Legendre polynomial to the angular distributions for emission energies with data. The coefficients of the Legendre polynomial for intermediate emission energies without data can then be obtained by linear interpolation.

The important difference between such a conventional fit and GP regression is the number of basis functions. Whereas their number is fixed and limited in a conventional fit, GP regression amounts to a fit with an infinite number of basis functions [1, Sect. 2.2]. The length scales regulate the number of basis functions that effectively contribute to the solution. Hyperparameter optimization decreases the length scales for unpredictable data which leads to a greater number of contributing basis functions and consequently to greater flexibility and larger uncertainties in the predictions. This feature sets GP regression apart from a standard χ^2-fit. In the latter, the uncertainty of the solution depends to a much lesser extent on the (un)predictability of the data and much more on the number of data points and the fixed number of basis functions.

One truly novel element in the approach is the inclusion of the mass and charge number, which enables predictions for isotopes without data. We can easily imagine that different isotopes differ significantly by their physical properties. From this perspective, the idea to extrapolate the model bias to other isotopes should be met with healthy skepticism.

To get a first grasp on the validity of isotope extrapolations, let us consider again the hyperparameter optimization discussed in Section 3.3. The hyperparameters were adjusted on the basis of the isotopes in Table 1. These data are spread out over the periodic table and cover a range from carbon to thorium. In these data, similar trends of the model bias persist across significant ranges of the mass and charge number, which was the reason that the associated length scales retained high values during optimization. For instance, the experimental data of carbon and indium in Figure 2 show comparable structures of the model bias despite their mass differences.

However, the isotopes considered in the optimization are not very exotic and gaps between them are at times large. Further, these isotopes are certainly not a random sample taken from the periodic table and therefore most theoretical guarantees coming from estimation theory do not hold. So how confident can we be about isotope extrapolation?

To provide a basis for the discussion of this question, Figure 2 also contains predictions for oxygen and cadmium. Importantly, the associated experimental data have not been used for the hyperparameter optimization and in the GP regression. The predictions follow well the trend of the experimental data. The supposedly 2σ uncertainty bands, however, include less than 95% of the data points. This observation points out that there are systematic differences between isotopes. Therefore, due to the sample of isotopes not being a random sample, uncertainty bands should be interpreted with caution.
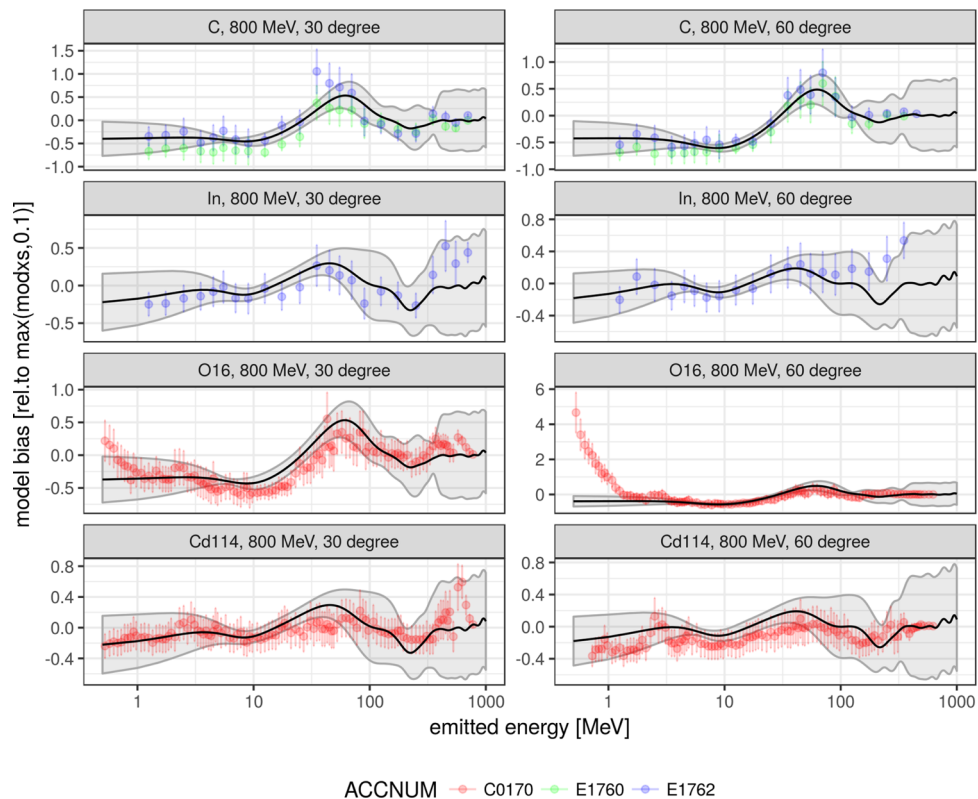
Fig. 2. Model bias of INCL in terms of equation (30) in the $(p,X)n$ double differential spectra for 800 MeV incident protons and different isotopes as predicted by GP regression. A missing mass number behind the isotope symbol indicates natural composition. The uncertainty band of the prediction and the error bars of the experimental data denote the 2σ confidence interval. Carbon and indium were taken into account in the GP regression but not cadmium and oxygen. The experimental data are colored according to the associated access number (ACCNUM) in the EXFOR database. This shows that all displayed data come from just three experiments [20–22].

One way to alleviate this issue could be to add a covariance function to equation (29) which only correlates spectrum values of the same isotope but not between isotopes. This measure would lead to an additional uncertainty component for each isotope, which only decreases if associated data are included in the GP regression.

A very extreme mismatch up to 500% between the prediction and experimental data occurs for oxygen at 60° and emission energies below 1 MeV. The creation of low energy neutrons is governed by de-excitation processes of the nucleus suggesting that the angular distribution of emitted particles is isotropic. This property holds for the model predictions but not for the experimental data. The experimental data exhibits a peak at 60° which is about a factor five higher than at 30°, 120° and 150°. The origin of this peculiarity may deserve investigation but is outside the scope of this work. For the sake of argument I assume that it is indeed a reflection of some property of the nucleus.

Because the data in Table 1 do not include any measurements below 1 MeV emission energy in this mass range, both hyperparameter optimization and GP regression had no chance to be informed about such a feature. In this case, the predictions and uncertainties are determined by nearby values or – if there are not any – by the prior expectation.

If we had been aware of such effects, we could have used it as a component in the covariance function to reflect our large uncertainty about the spectrum for low emission energies. Otherwise, the only sensible data-driven way to provide predictions and uncertainties for unobserved domains is to assume that effects are similar to those in some observed domains. Of course, it is our decision which domains are considered similar. The employed covariance function in equation (29) incorporates the assumption that nearby domains in terms of mass charge, angle, etc. are similar. However, we are free to use any other assumption about similarity to construct the covariance function. As a side note, also results of other models relying on different assumptions could serve as a basis for uncertainties in unobserved regions. Such information can be included in GP regression in a principled way.

4 Summary and outlook

Sparse GP regression, a non-parametric estimation technique of statistics, was employed to estimate the model bias of the C++ version of the Liège INCL coupled to the evaporation code ABLA07. Specifically, the model bias in the prediction of double-differential inclusive neutron spectra over the complete nuclide chart was investigated for incident

protons above 100 MeV. Experimental data from the EXFOR database served as the basis of this assessment. Roughly ten thousand data points were taken into account. The hyperparameter optimization was done on a computing cluster whereas the GP regression itself on a desktop computer. The obtained timings indicate that increasing the number of data points by a factor of ten could be feasible.

For this specific application, it was shown that GP regression produces reasonable results for isotopes that have been included in the procedure. It was argued that the validity of predictions and uncertainties for isotopes not used in the procedure depends on the validity of the assumptions made about similarity between isotopes. As a simple benchmark, the isotopes oxygen and cadmium, which have not been taken into account in the procedure, were compared to the respective predictions. The agreement between prediction and experimental data was reasonable but the 95% confidence band sometimes misleading and should therefore be interpreted with caution. Accepting the low energy peak of oxgyen at 60° in the data as physical reality, the low energy spectrum of oxygen served as an example where the similarity assumption between isotopes did not hold.

As for any other uncertainty quantification method, it is a hard if not impossible task to properly take into account unobserved phenomena without any systematic relationship to observed ones. The existence of such phenomena are *unknown unknowns*, their observation is tagged shock, surprise or discovery, and they potentially have a huge impact where they appear.

Even though GP regression cannot solve the philosophical problem associated with unknown unknowns, knowledge coming from the observation of new effects can be taken into account by modifying the covariance function. For instance, the peculiar peak in the oxygen spectrum suggests the introduction of a term in the covariance function which increases the uncertainty of the spectrum values in the low emission energy domain. The ability to counter new observations with clearly interpretable modifications of the covariance function represents a principled and transparent way of knowledge acquisition.

The scalability and the possibility to incorporate prior assumptions by modeling the covariance function makes sparse GP regression a promising candidate for global assessments of nuclear models and to quantify their uncertainties. Because the formulas of GP regression are structurally identical to those of many existing evaluation methods, GP regression should be regarded more as a complement than a substitute, which can be interfaced with existing evaluation methods.

References

1. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, Mass., 2006)
2. D.W. Muir, A. Trkov, I. Kodeli, R. Capote, V. Zerkin, *The Global Assessment of Nuclear Data, GANDR* (EDP Sciences, 2007)
3. H. Leeb, D. Neudecker, T. Srdinko, Nucl. Data Sheets **109**, 2762 (2008)
4. M. Herman, M. Pigni, P. Oblozinsky, S. Mughabghab, C. Mattoon, R. Capote, Y.S. Cho, A. Trkov, Tech. Rep. BNL-81624-2008-C P, Brookhaven National Laboratory, 2008
5. J. Quiñonero-Candela, C.E. Rasmussen, J. Mach. Learn. Res. **6**, 1939 (2005)
6. E. Snelson, Z. Ghahramani, Sparse Gaussian Processes Using Pseudo-Inputs, in *Advances in Neural Information Processing Systems* (2006), pp. 1257–1264
7. D. Mancusi, A. Boudard, J. Cugnon, J.C. David, P. Kaitaniemi, S. Leray, *New C++ Version of the Liège Intranuclear Cascade Model in Geant4* (EDP Sciences, 2014), p. 05209
8. D. Mancusi, A. Boudard, J. Cugnon, J.C. David, P. Kaitaniemi, S. Leray, Phys. Rev. C **90**, 054602 (2014)
9. A. Kelic, M.V. Ricciardi, K.-H. Schmidt, in *Proceedings of the Joint ICTP-IAEA Advanced Workshop on Model Codes for Spallation Reactions* (ICTP Trieste, Italy, 2008)
10. N. Otuka, E. Dupont, V. Semkova, B. Pritychenko, A. Blokhin, M. Aikawa, S. Babykina, M. Bossant, G. Chen, S. Dunaeva et al., Nucl. Data Sheets **120**, 272 (2014)
11. B.J.N. Blight, L. Ott, Biometrika **62**, 79 (1975)
12. M.T. Pigni, H. Leeb, Uncertainty estimates of evaluated 56fe cross sections based on extensive modelling at energies beyond 20 MeV, in *Proc. Int. Workshop on Nuclear Data for the Transmutation of Nuclear Waste. GSI-Darmstadt, Germany* (2003)
13. G. Schnabel, Ph.D. thesis, Technische Universität Wien, Vienna, 2015
14. G. Schnabel, H. Leeb, EPJ Web Conf. **111**, 09001 (2016)
15. R.H. Byrd, P. Lu, J. Nocedal, C. Zhu, SIAM J. Sci. Comput. **16**, 1190 (1995)
16. *IAEA benchmark 2010*, https://www-nds.iaea.org/spallations/
17. J.C. David, Eur. Phys. J. A **51**, 68 (2015)
18. MYRRHA: An innovative research installation, http://sckcen.be/en/Technology_future/MYRRHA
19. R Development Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2008)
20. W.B. Amian, R.C. Byrd, C.A. Goulding, M.M. Meier, G.L. Morgan, C.E. Moss, D.A. Clark, Nucl. Sci. Eng. **112**, 78 (1992)
21. T. Nakamoto, K. Ishibashi, N. Matsufuji, N. Shigyo, K. Maehata, S.I. Meigo, H. Takada, S. Chiba, M. Numajiri, T. Nakamura et al., J. Nucl. Sci. Technol. **32**, 827 (1995)
22. K. Ishibashi, H. Takada, T. Nakamoto, N. Shigyo, K. Maehata, N. Matsufuji, S.I. Meigo, S. Chiba, M. Numajiri, Y. Watanabe et al., J. Nucl. Sci. Technol. **34**, 529 (1997)