
UPSALA
UNIVERSITET

The nuclear data evaluation pipeline of Uppsala University (NEPU)

– addressing model defects and data inconsistencies

Alf Göök, Henrik Sjöstrand & Erik Andersson Sundén [1]
Uppsala University

2023-12-18

[1] with inspiration and previous work from:
G. Schnabel, P. Helgesson and J. Hansson

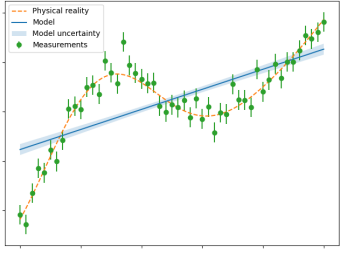
UPPSALA UNIVERSITET

Model Defects

A model that cannot reproduce the underlying truth, no matter its parameters, can have sever consequences

- Evaluation become biased towards the model.[2,3]
- Uncertainties will be underestimated – often severely.[2,3]

$\rightarrow$ Model defects must be addressed.



"All models are wrong,
some are useful."
– G.E. Box, 1976

[2] P. Helgesson, Ph.D. thesis, Uppsala University (2018)

[3] G. Schnabel, Ph.D. thesis, TU Wien (2015)

Background

The goal of this project is to develop a pipeline for nuclear data evaluation, that implements (and further develops) methodology to treat **model defects** and **inconsistent experimental data** that has originated in research activities at UU. In addition, the pipeline should

- **automatize** as much as possible the steps involved in an evaluation
- create fully **reproducible ND evaluations**
- provide an **intuitive framework** for ND evaluation

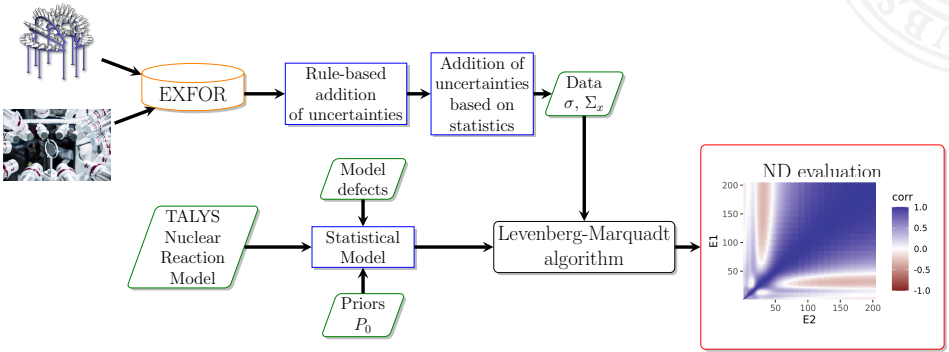
Nuclear Data Sheets 173 (2021) 239–284

Conception and Software Implementation of a Nuclear Data Evaluation Pipeline

G. Schnabel,[1,*] H. Sjöstrand,[2] J. Hansson,[2] D. Rochman,[3] A. Koning,[1] and R. Capote[1]

The ND Evaluation Pipeline

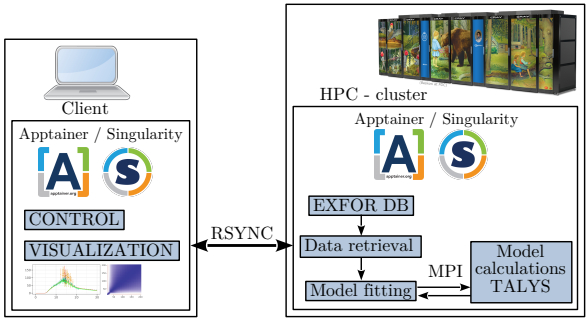
The pipeline is a bundle of software packages, along with a set of scripts that each perform a step in a nuclear data evaluation



Underlying assumption: knowledge about the cross-sections can be represented by a multi-variate normal distribution of TALYS parameters

High Performance Computing

- Pipeline can run entirely on HPC cluster
- MPI wrapper for Talys
  – large scale parallelization
- Currently running on UPPMAX[a]
  – Rackham cluster
- A full evaluation, including the generation of random files, can be performed in a few hours
- Greatly facilitates the testing and validation of ideas and methods



[a]Uppsala Multidisciplinary Center for Advanced Computational Science

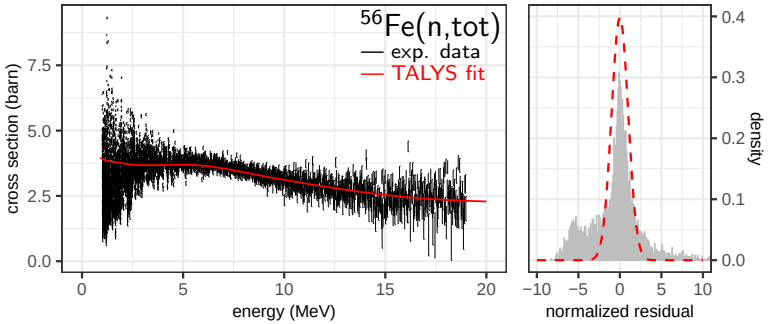
Workflow in the Pipeline

is divided into a number of steps, each represented by an R-script.

1. Data is retrieved from EXFOR $\rightarrow$ Mapped to TALYS predictions
2. Rule-based correction of uncertainties in data
3. Correction of uncertainties based on statistics
4. Talys parameter sensitivity evaluation
5. Setup of GP for energy dependence of parameters
6. Parameter optimization using the LM algorithm
7. Setup of GP in the observable domain
8. Re-optimization using the LM algorithm
9. Calculation of MVN approximation of the posterior pdf
10. Generation of random files

Treatment of Random Uncertainties

TALYS models the average cross-section, while experiments observe (unresolved) resonance structure. Therefore, the variance of the data around the mean cross-section is much larger than their reported random (statistical) variance.



Not treating this **model defect** can lead to biased results and strongly underestimated uncertainties of the fit.

Treatment of Random Uncertainties

- Estimate the distribution of data around a smooth energy-averaged cross-section.
- The smooth cross-section should be influenced by the model as little as possible.
- Assuming that the distribution is Normal, a Gaussian Process (GP) seems like a good candidate.
- A GP models data by estimating the correlation between close-lying points with a covariance function, e.g.

$$\mathrm{cov}(y(x_i), y(x_j)) = \sigma^2 \exp\left[-\frac{(x_i - x_j)^2}{2\lambda^2}\right] + \tau^2 \delta_{ij}$$

- The hyper-parameter λ controls the length-scale (smoothness)
- The random error in the data is modeled by the *nugget parameter* τ

Treatment of Random Uncertainties

To determine the distribution of data around a smooth mean function, we model it using a *heteroscedastic GP* [4].

$$\text{cov}(y(x_i), y(x_j)) = \sigma^2 \exp\left[-\frac{(x_i - x_j)^2}{2\lambda^2}\right] + \tau^2 \delta_{ij}$$

- The *heteroscedastic GP* introduces latent variance variables, placed under a GP to allow a smoothly varying *nugget parameter* – variance of data around the mean

$$\tau^2 \delta_{ij} \rightarrow \delta(x_1), \delta(x_2), ..., \delta(x_n), \quad \delta(x) \sim GP$$

[4] https://CRAN.R-project.org/package=hetGP

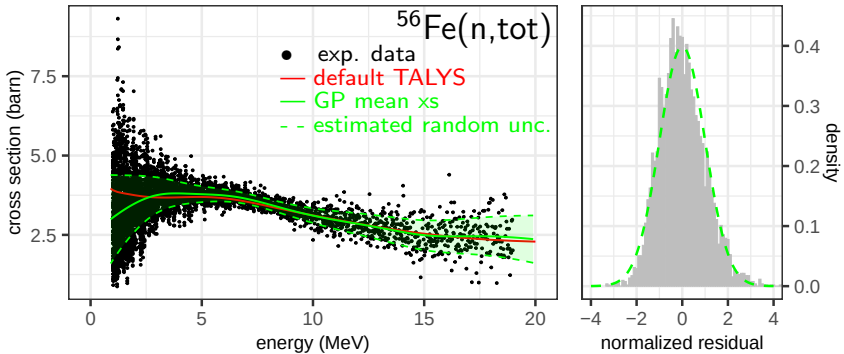
Treatment of Random Uncertainties

In practice...

- To separate random and systematic uncertainties, we apply the procedure experiment by experiment.
- Experiments for which to apply the procedure are selected based on the energy resolution in the experiment.
- The length scale is determined from a Marginal Likelihood Optimization on default TALYS predictions.
- Hyper-parameters for the heteroscedastic GP are optimized with the lengthscale λ fixed – simultaneous inference of mean xs and the variance of the data

Treatment of Random Uncertainties

Example application on ^{56}Fe(n,tot) data[5]

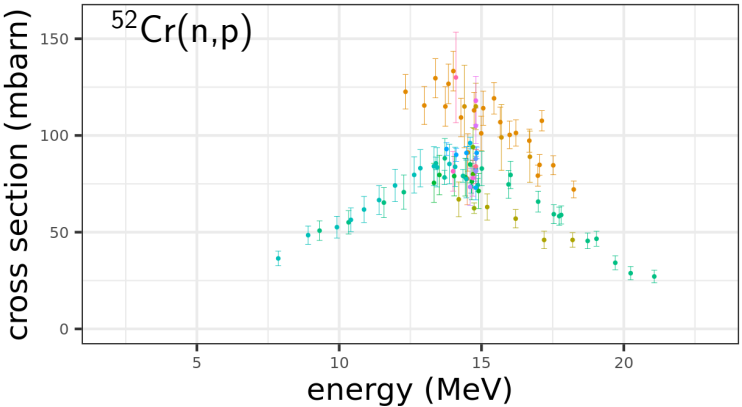


Finally, the reported random uncertainties of the selected experiments are replaced by those estimated by the heteroscedastic GP before fitting TALYS parameters.

[5] E.Cornelis,L.Mewissen,F.Poortmans – EXFOR entry 22316001

Inconsistent data

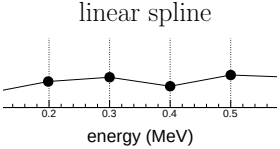
What to do when the difference between the datasets is too large to be explained by the reported uncertainties?



Correction of Systematic Uncertainties

The presence of unrecognized uncertainties can be identified if data-sets are inconsistent with each other.

- A linear spline with a MVN prior on the values at the knot-points – induces a Gaussian process
- We use a prior on the second derivative at the knot-points



$$\nabla_i^2 = \left(\frac{\partial^2 y}{\partial E^2}\right)_i$$

$$\nabla^2 \sim \mathcal{N}(\mu, \Sigma)$$

- The prior is based on the default TALYS calculation
  - μ – default TALYS prediction
  - Σ – diag(μ)

Correction of Systematic Uncertainties

The presence of unrecognized uncertainties can be identified if data-sets are inconsistent with each other.

- The resulting distribution of linear splines is used in a marginal likelihood optimization of normalization uncertainties in the experiments.

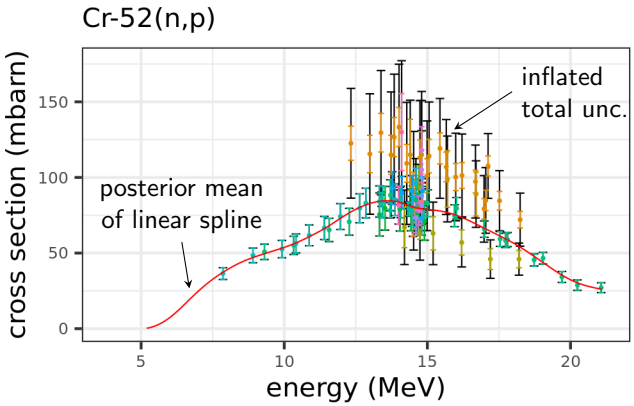
$$\det(\mathbf{\Sigma})^{-1/2} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T \mathbf{\Sigma} (\vec{x}-\vec{\mu})\right]$$

$$\vec{x} = \text{experimental cross sections}$$
$$\vec{\mu} = \text{linear spline}$$
$$\mathbf{\Sigma} = \mathbf{\Sigma}_{\text{exp}} + \mathbf{\Sigma}_{\text{spline}} + \mathbf{\Sigma}_{\text{extra}}$$

Added normalization uncertainty of each experiment is contained in $\mathbf{\Sigma}_{\text{extra}}$

Correction of Systematic Uncertainties

Outliers are assigned inflated normalization uncertainty.



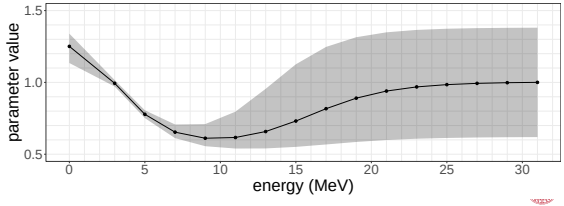
Treatment of Model Defects

- GP in the parameter domain[6]

- Incident energy-dependent variation of parameters around *global* value
- The variation is modeled with a Gaussian process (GP)

$$\sigma(E_j) = f(E_j; \vec{p} + \vec{\delta}(E_j)) + \varepsilon, \qquad \vec{\delta}(E) \sim \text{GP}$$

- Smooth variation of parameter values with energy
- Consistent physics description at each energy
- TALYS conserves the sum-rules
- In the presence of model defects, parameter uncertainty is mainly constrained were there is data

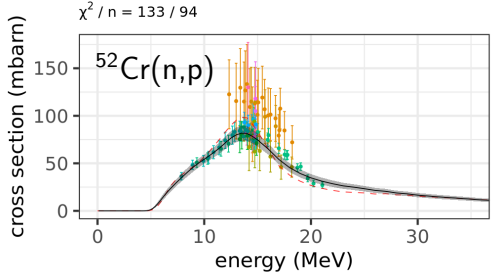


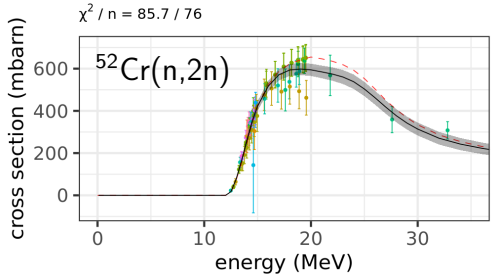
[6] P. Helgesson & H. Sjöstrand, Ann. Nucl. Energy 120 (2018) 35-47

EXAMPLE OF RESULTS

APPLICATION OF THE PIPELINE ON ^{52}CR CROSS-SECTIONS

Exclusive particle production xs

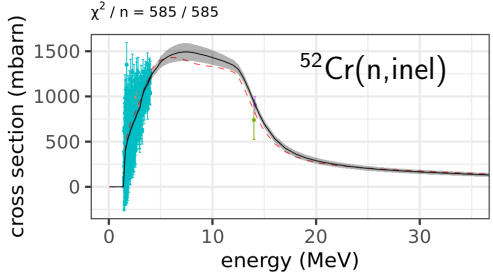


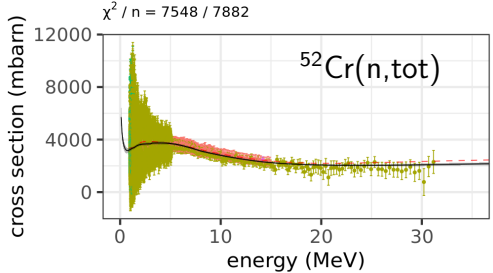


— posterior mean (black)

- - prior mean (red)

Channels affected by resonance structure





— posterior mean (black)

- - prior mean (red)

- Low χ^2 for the (n,tot)-channel could indicate an overestimated random uncertainty by the heteroscedastic GP

Cross-validation

number of degrees of freedom is not well defined for non-linear models with priors

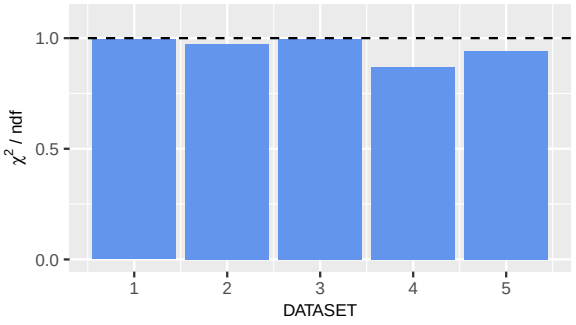
- 5-fold cross-validation
  - Experimental data is randomly divided into 5 subsets
  - each time 20% is left out
- The pipeline is executed in full on each data-set
- χ^2 for the data left out from the fit

$$\chi^2 = r^T \mathbf{\Sigma}^{-1} r$$
$$\mathbf{\Sigma} = \mathbf{\Sigma}_{\text{exp}} + \mathbf{\Sigma}_{\text{res}} + \mathbf{\Sigma}_{\text{extra}} + \mathbf{\Sigma}_{\text{GP}}$$

$\mathbf{\Sigma}_{\text{exp}} =$ experimental covariance matrix
$\mathbf{\Sigma}_{\text{res}} =$ estimated random unc. due to resonance structure
$\mathbf{\Sigma}_{\text{extra}} =$ added normalization unc. through MLO
$\mathbf{\Sigma}_{\text{GP}} =$ cov. function of the residual model defect GP

Cross-validation – results

- Overall good performance.
- Indicates that the automated procedures perform well.
- A slight tendency to overestimate uncertainties is noted.

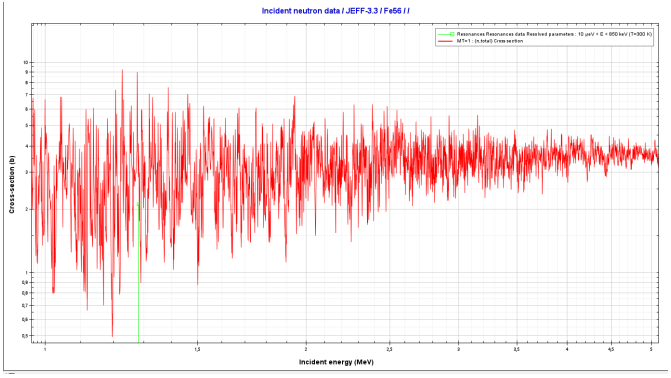


DATASET	ndf	χ^2	χ^2/ndf
1	1731	1735.56	1.00
2	1731	1732.94	1.00
3	1729	1755.52	1.02
4	1730	1498.83	0.87
5	1730	1659.85	0.96
sum	8651	8378.70	0.97±0.02

Outlook: Low-energy structure

Energy-dependent Talys parameters cannot reproduce rapid variation in low energy region (1-5 MeV).

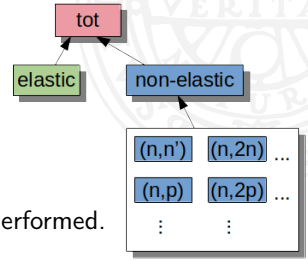
The evaluation must respect the sum-rules.



UPPSALA UNIVERSITET

Construction of a Defect model

Sum-rules are defined in accordance with TALYS notation



$$\sigma_{\text{tot}} = \sigma_{\text{el}} + \sigma_{\text{non-el}}$$

$$\sigma_{\text{non-el}} = \sum_{n}\sum_{p}\sum_{d}\sum_{t}\sum_{h}\sum_{\alpha} \sigma^{\text{ex}}(n, p, d, t, h, \alpha)$$

- An energy grid is chosen, on which the evaluation is to be performed.
  - Intermediate energies are linearly interpolated.
  - At each energy one 'parameter' per **open** exclusive cross-section is introduced.
- From the parameters other cross-sections can be calculated, for example

$$\sigma_{(\text{n,xn})} = \sum_{n}\sum_{p}\sum_{d}\sum_{t}\sum_{h}\sum_{\alpha} n\,\sigma^{\text{ex}}(n, p, d, t, h, \alpha)$$

Using the Defect model

The sum-rules constitutes a linear model, whose output is fully represented by

$$\varepsilon_m = \mathbf{J}\beta$$

where β is the vector of exclusive cross-sections on the chosen energy grid.

This simple construction is intended to be used to model the residual of the TALYS fit, i.e. the data vector is modeled

$$\mathbf{D} = \mathbf{f}(\mathbf{x};\mathbf{p}) + \varepsilon + \varepsilon_m,$$

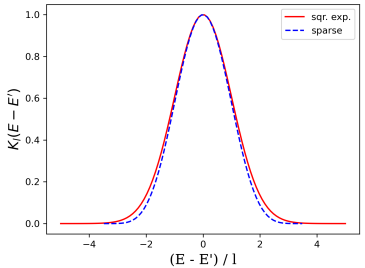
where $\mathbf{f}$ is the TALYS output,
ε is the experimental uncertainty and
ε_m is the defect model

Gaussian Process on the Defect model

The vector β is then placed under a Multi-variate normal prior.

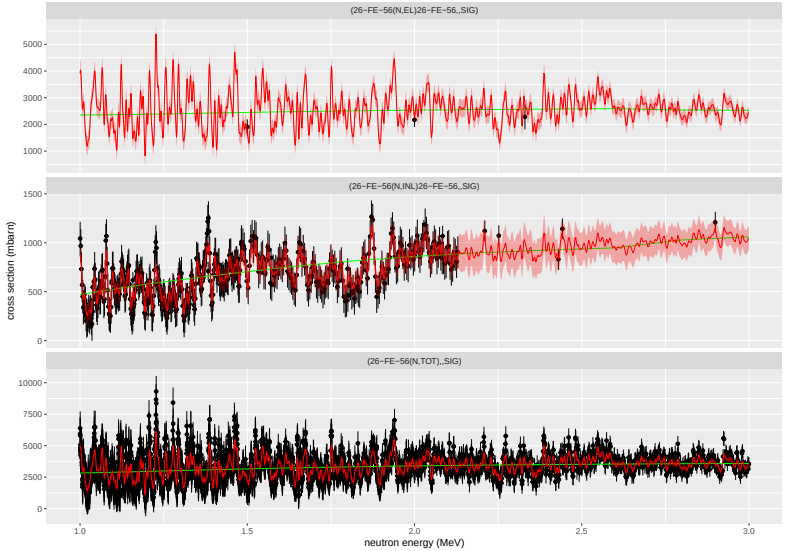
$$\beta \sim \mathcal{N}(0, \mathbf{K}) \quad \{\mathbf{K}\}_{i,j} = \nu K_l(E_i, E_j)$$

- Exclusive cross-sections are considered uncorrelated.
- Within each exclusive channel, prior correlation is parameterized based on the energy distance by a covariance function $K_l(E_i, E_j)$.



In other words: Energy dependence of the exclusive cross-sections is modeled using a Gaussian Process.

A first test... on ^{56}Fe



UPPSALA UNIVERSITET

Summary

- We are developing a pipeline for evaluation of the fast energy range
  - based around the TALYS code system
  - implements automated procedures for
    - treatment of inconsistent experimental data
    - treatment of model defects
    - new treatment of resonance structure based on heteroscedastic GP
  - designed for fully reproducible ND evaluations
  - capability to take advantage of large-scale parallel computing
- Application of the pipeline on ^{52}Cr has been presented
  - cross-validation shows that the model, the automated correction procedures, and the treatment of model defects work well

Thank's for the attention!

Gaussian Process on the Defect model

The covariance matrix that represents the MVN prior on the model parameters is block diagonal, with one block per open exclusive channel

$$\mathbf{K} = \begin{bmatrix} \mathbf{K_1} & & \\ & \ddots & \\ & & \mathbf{K_k} \end{bmatrix}$$

The model parameters and experimental data are then joint normally distributed. The full covariance matrix consists of four blocks

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{K} & \mathbf{K}\mathbf{J}^T \\ \mathbf{J}\mathbf{K} & \mathbf{J}\mathbf{K}\mathbf{J}^T + \Sigma_{\text{exp}} \end{bmatrix}.$$

By conditioning on the observed residual $\mathbf{r}$ we can find the conditional mean and covariance of the parameter vector.

$$\overline{\beta} = \mathbf{\Sigma}_{12}\,\mathbf{\Sigma}_{22}^{-1}\,\mathbf{r}, \quad \overline{\mathbf{\Sigma}}_{11} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}$$