UPPSALA UNIVERSITET

A Nuclear Data Evaluation Pipeline for the Fast Neutron Energy Range

–using heteroscedastic Gaussian processes to treat model defects

Alf Göök, Henrik Sjöstrand, Erik Andersson Sundén, Joachim Hansson
Uppsala University

WONDER-2023 || 2023-07-08

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]

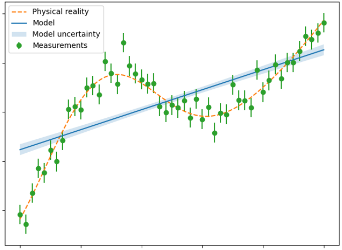
UPPSALA UNIVERSITET

Model Defects

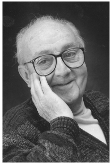
A model that cannot reproduce the underlying truth, no matter its parameters, can have sever consequences

- Evaluation become biased towards the model.[1,2]
- Uncertainties will be underestimated – often severely.[1,2]

→ Model defects must be addressed.



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

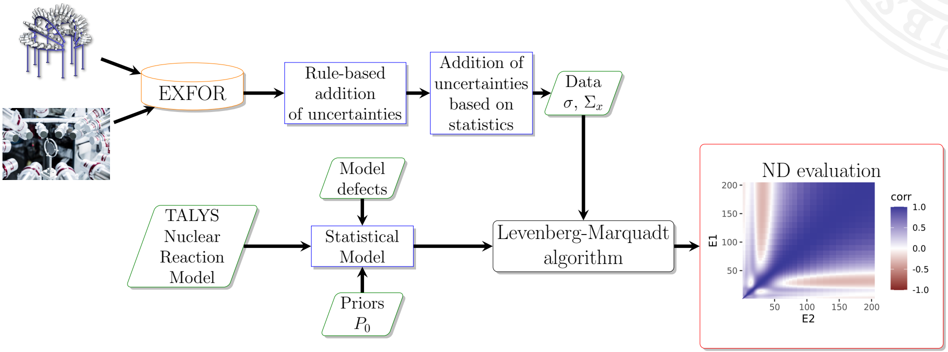


[1] P. Helgesson, Ph.D. thesis, Uppsala University (2018)

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

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

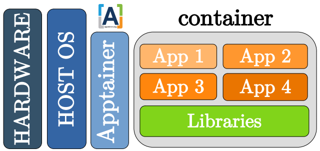


Underlying assumption: knowledge about the cross-sections can be represented by a multi-variate normal distribution of TALYS parameters

UPPSALA UNIVERSITET

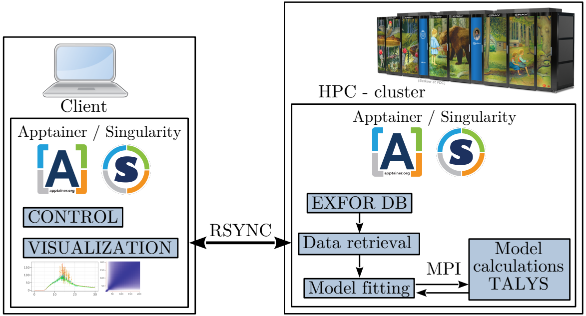
Containerization

- To facilitate **reproducibility** the pipeline is containerized.
- A container is a standard unit of software that packages up code and all its dependencies so an application runs quickly and reliably from one computing environment to another – bypassing complex installation procedures.
- This allows the evaluation to be archived and distributed for others to replicate, no matter what version of Linux they are running.
- We use the container platform Apptainer, created specifically to run on high-performance computing (HPC) clusters.



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

UPPSALA
UNIVERSITET

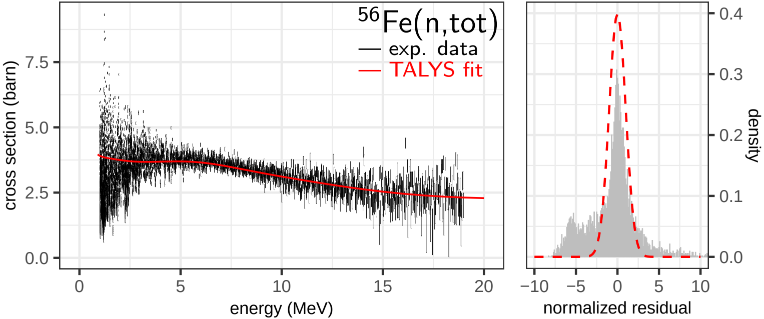
Workflow in the Pipeline

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.

UPPSALA UNIVERSITET

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

UPPSALA UNIVERSITET

Treatment of Random Uncertainties

To determine the distribution of data around a smooth mean function, we model it using a *heteroscedastic GP* [3].

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

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

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

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

UPPSALA UNIVERSITET

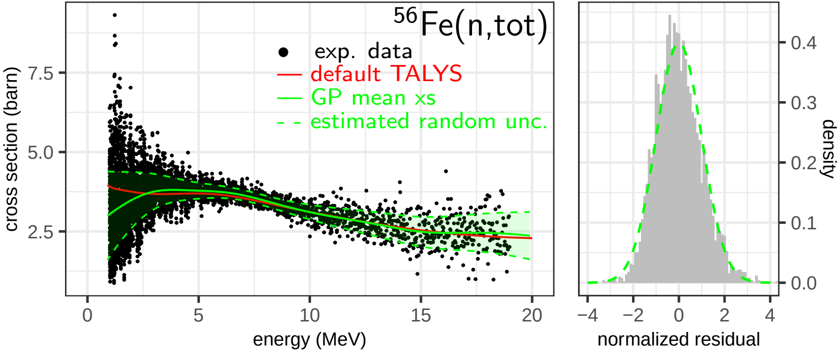
Treatment of Random Uncertainties

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[4]



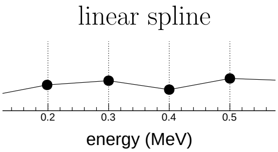
Finally, the reported random uncertainties of the selected experiments are replaced by those estimated by the heteroscedastic GP before fitting TALYS parameters.

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

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)

UPPSALA UNIVERSITET

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.

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

$$\vec{x} = \text{experimental cross sections}$$

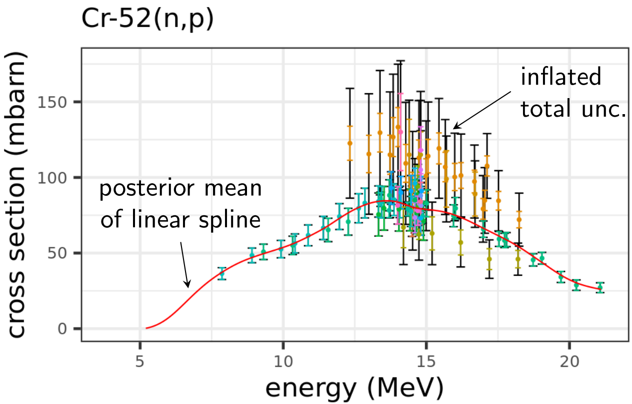
$$\vec{\mu} = \text{linear spline}$$

$$\mathbf{\Sigma} = \mathbf{\Sigma}_{\text{exp}} + \mathbf{\Sigma}_{\text{spline}} + \mathbf{\Sigma}_{\text{extra}}$$

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

Correction of Systematic Uncertainties

Outliers are assigned inflated normalization uncertainty.



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

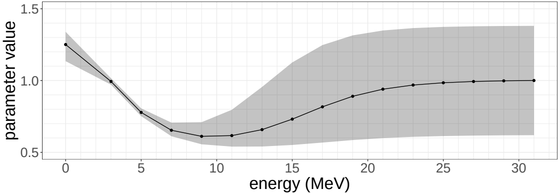
Treatment of model defects

Treatment of Model Defects
- GP in the parameter domain[5]

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

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

- 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



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

UPPSALA UNIVERSITET

Treatment of Model Defects
- GP in the cross-section domain

- An advantage of the parameter GP is the conservation of the physics
- The freedom may not be large enough to describe all model defects

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

- ε_m - describes the residual deviation between experiments and model
- ε_m - is not added to the final result (to not mess with the physics)
- The covariance structure of the GP is retained when fitting the model

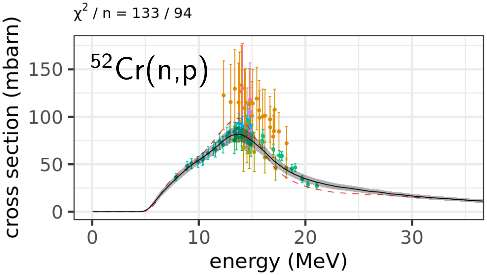
$$\text{cov}(\sigma(E_i), \sigma(E_j)) = A^2 \exp\left[-\frac{(E_i - E_j)^2}{2\lambda^2}\right]$$

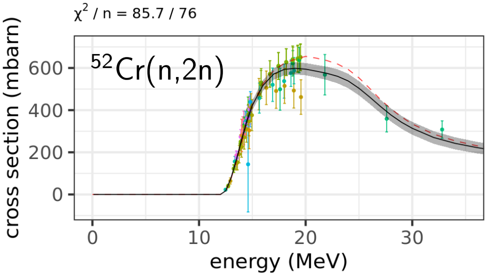
- representing the expected deviation of the model from the data
- Hyper-parameters (A, λ) are found using marginal likelihood optimization

Example of results

Application of the pipeline on
^{52}Cr cross-sections

Exclusive particle production xs

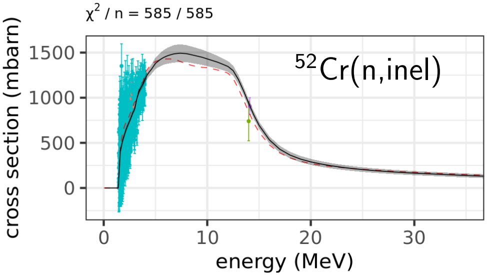


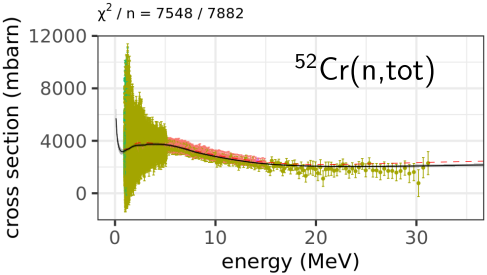


— posterior mean (black)

- - 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

UPPSALA UNIVERSITET

Cross-validation

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

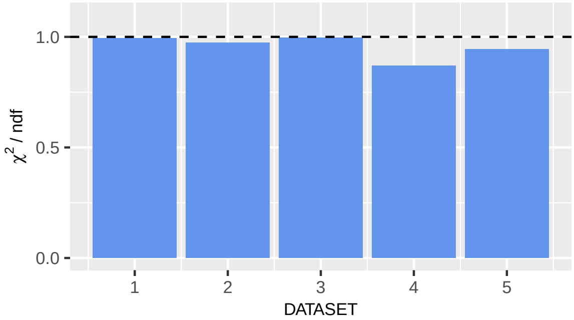
- 5-fold cross-validation
  - Experimental data is randomly divided into 5 subsets
  - each time 20% is left out
- The pipeline is executed in full on each data-set
- χ^2 for the data left out from the fit

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

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

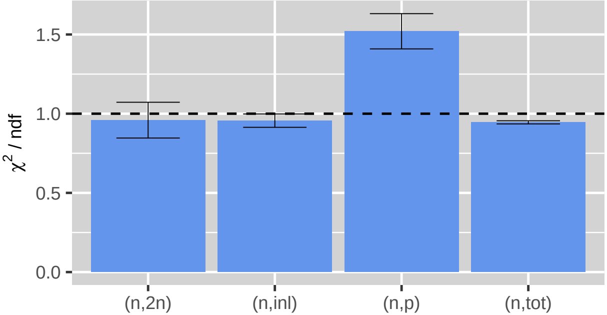
Cross-validation – results

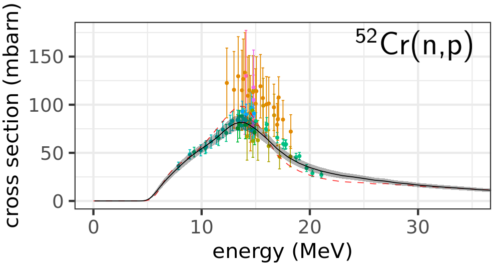
- Overall good performance.
- Indicates that the automated procedures perform well.
- A slight tendency to overestimate uncertainties is noted.



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

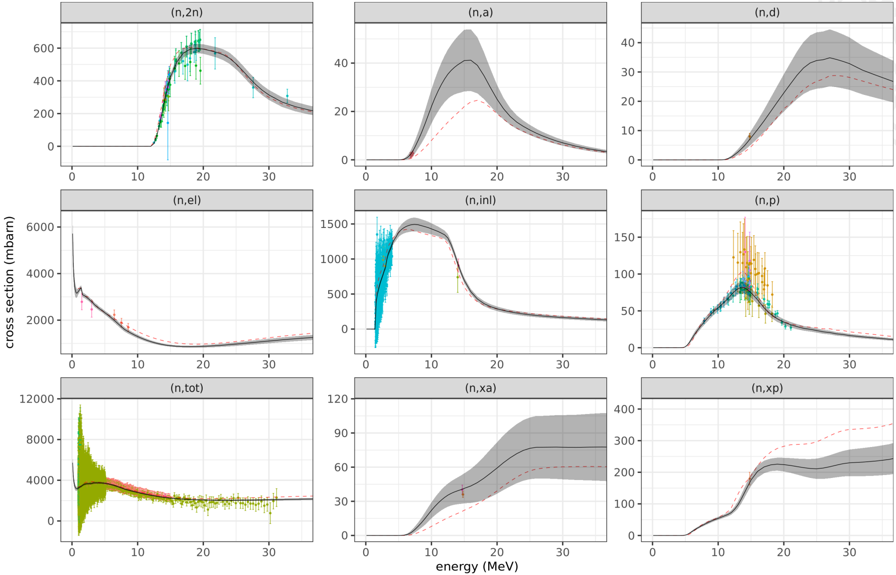
Cross-validation – results per channel





- Only channels with > 5 data points included here.

Results – all fitted channels



UPPSALA UNIVERSITET

Summary & Conclusion

- 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

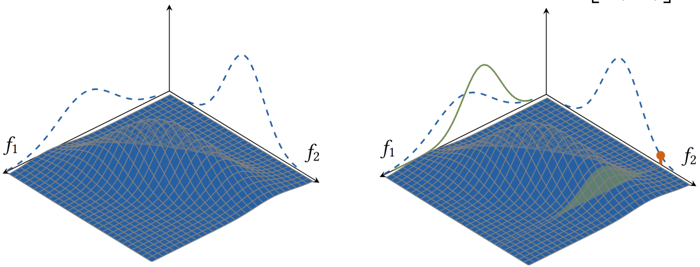
Thank you for your attention!




What's a Gaussian Process?

Assume we have cross-section values at two different energies: $\sigma(E_1)$,$\sigma(E_2)$

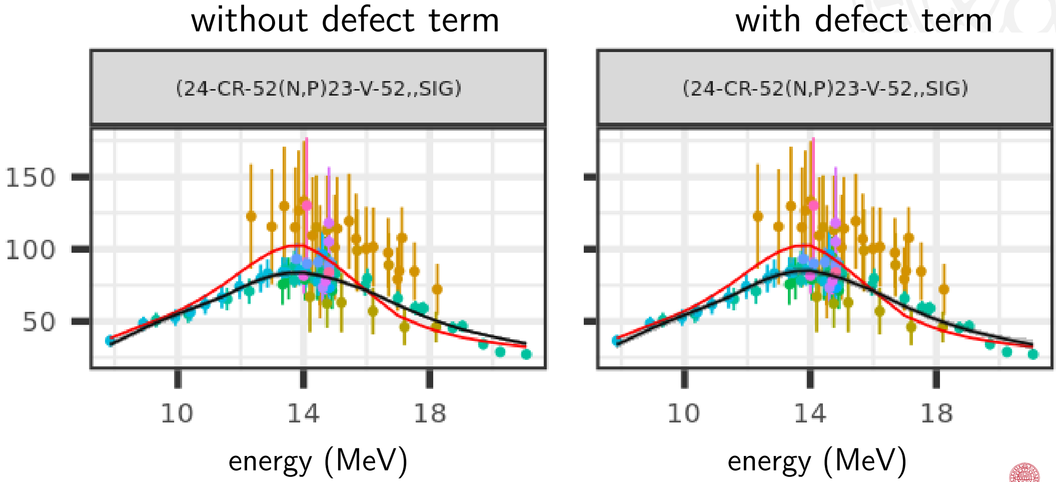
Assume that the values are joint normally distributed.
$$\begin{bmatrix}\sigma(E_1)\\ \sigma(E_2)\end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \Sigma\right)$$



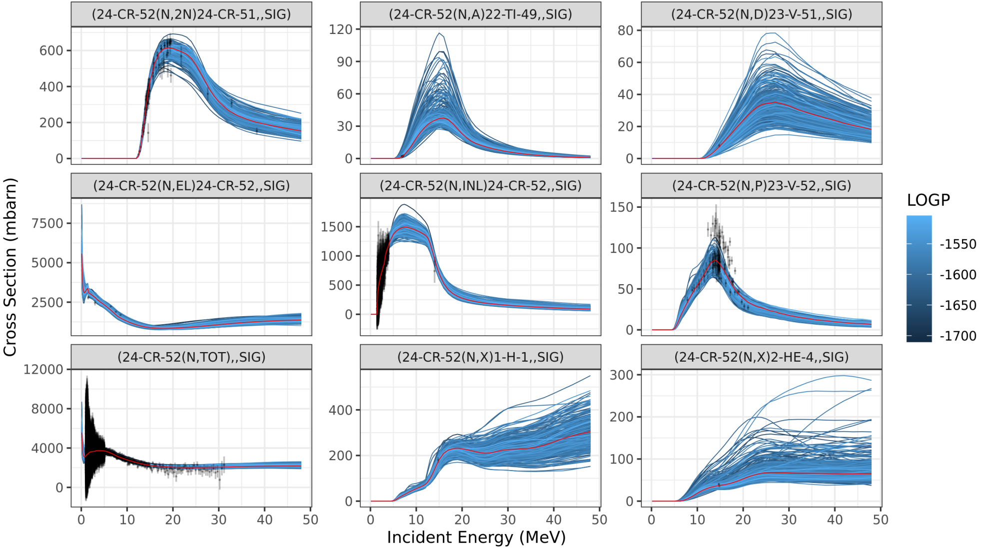
If we observe one of the values $\sigma(E_2) = f_2$
we get a conditional distribution for the other value.

This can be extended to larger dimensions and used as a regression model.

Effect of GP in observable



Cross section samples



UPPSALA UNIVERSITET