Applying Random Sampling methods to data analysis for uncertainty production, with an Open source and Open science outlook

Greg Henning

Applying Random Sampling methods to data analysis for uncertainty production, with an Open source and Open science outlook.

Habilitation à Diriger les Recherches defense

June 27th, 2024 – IPHC, Strasbourg, France

Greg Henning

Université de Strasbourg, Centre National de la Recherche Scientifique, IPHC UMR 7178, F-67000 Strasbourg, France

Outline

- Context: Nuclear Data for application
- Uncertainty and Covariance with Random sampling
- High quality nuclear data with Open Science methods
- Implementing a full Monte-Carlo analysis code
- New code, New tool, New method

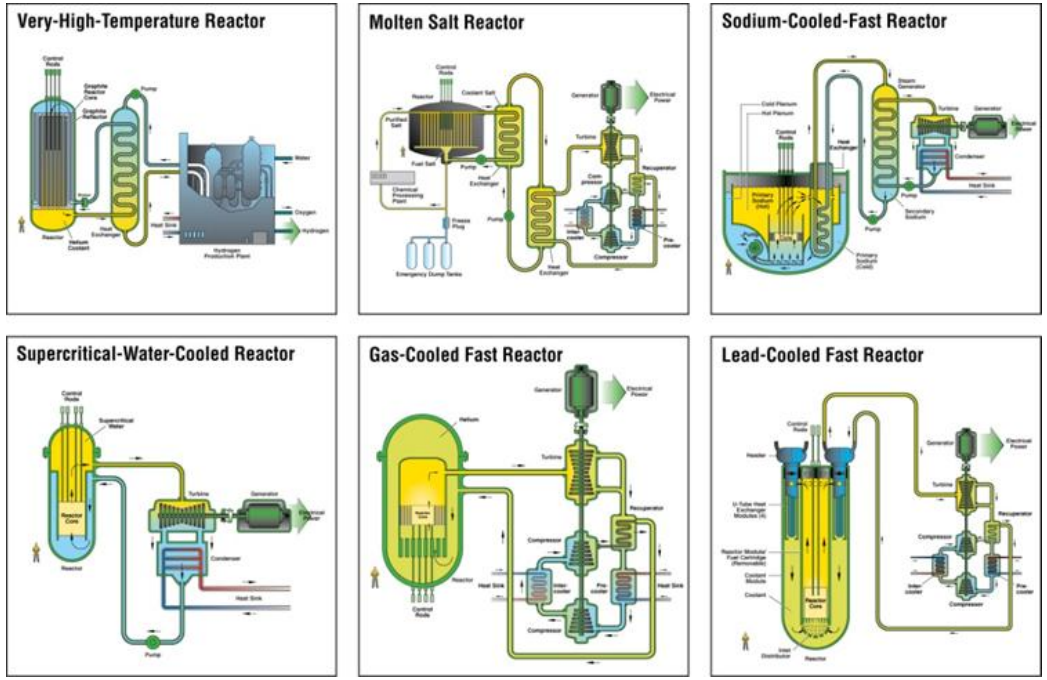
This work was funded in part by the CNRS multipartner NEEDS program, PACEN/GEDEPEON, and by the European Commission within the Sixth Framework Program through I3-EFNUDAT (EURATOM contract n°036434) and NUDAME (Contract n°FP6-516487), within the Seventh Framework Program through EUFRAT (EURATOM contract n°FP7-211499), through ANDES (EURATOM contract n°FP7-249671), from the Euratom research and training program 2014-2018 under grant agreement n°847594 (ARIEL) and under grant agreement n°847552 (SANDA).

Development of new reactors designs

Needs for future reactors

- Isotopic composition of fuel will be different (Th or Pu, instead of ^{235}U)
- Use of *fast* neutrons (compared to thermal).

- The current knowledge on thermal neutron induced reactions on the ^{235}U cycle isotopes of interest is not enough to characterize accurately fast neutron reactors, or breeding reactors.



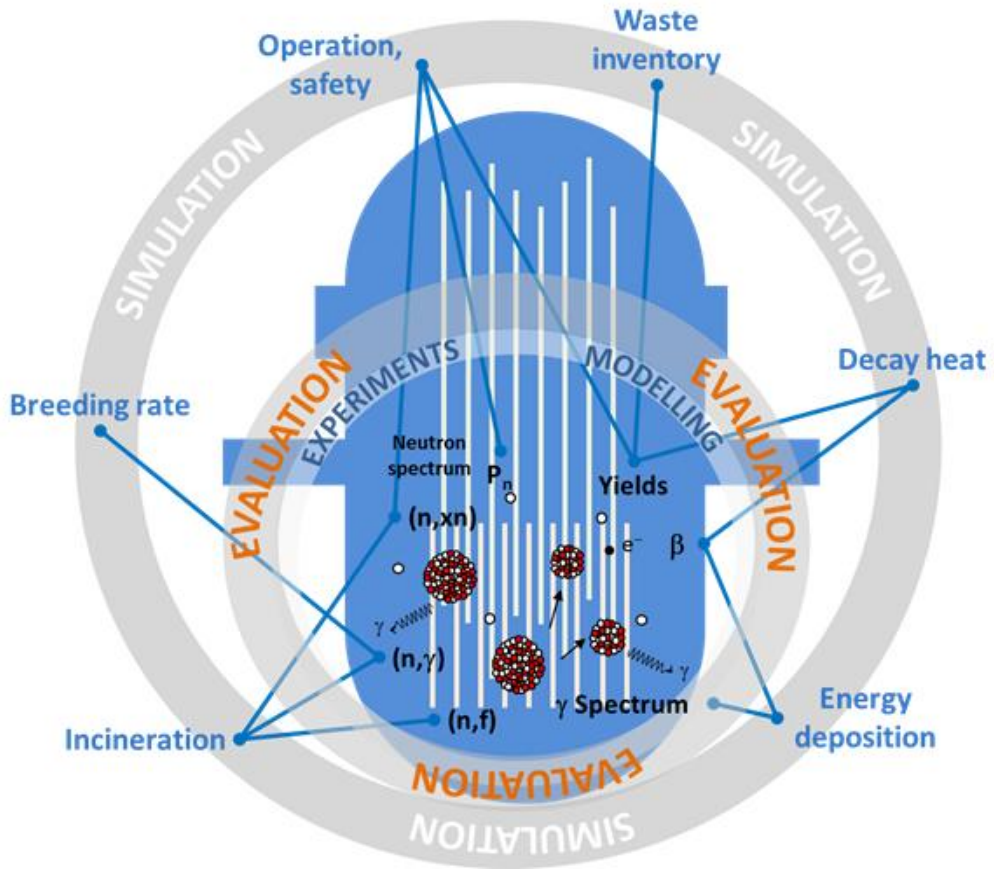
Development of new reactors designs

Needs for future reactors

- Numerical simulations done for the development of future reactor designs are using evaluated nuclear data (cross sections, energy distribution, yields, ..), compiled from microscopic measurements.

- « Evaluation » is the mathematical process that determines our best estimate of the nuclear data value from microscopic and integral measurements (including selection and normalization), and theoretical models.
- *Quality* of data (i.e. well documented uncertainty, covariance matrices, documentation of normalization factors and reference reactions used, …) plays an important role in the evaluation process.

- There is an important need for new and *complete* experimental data to improve evaluations.



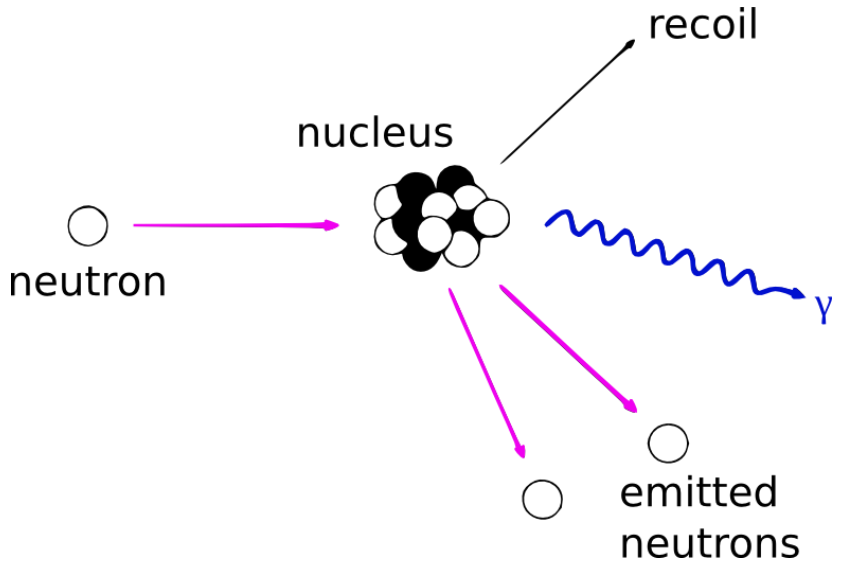
© M. Kerveno

Inelastic Neutron scattering

(n, xn) and (n, xn γ) reactions

- Energy loss mechanism for neutrons
- Production of γ rays
- Interaction by nuclear force only
- Modify neutron multiplicity and creates new isotopes

- Also, contribute to non-local effects in reactors.

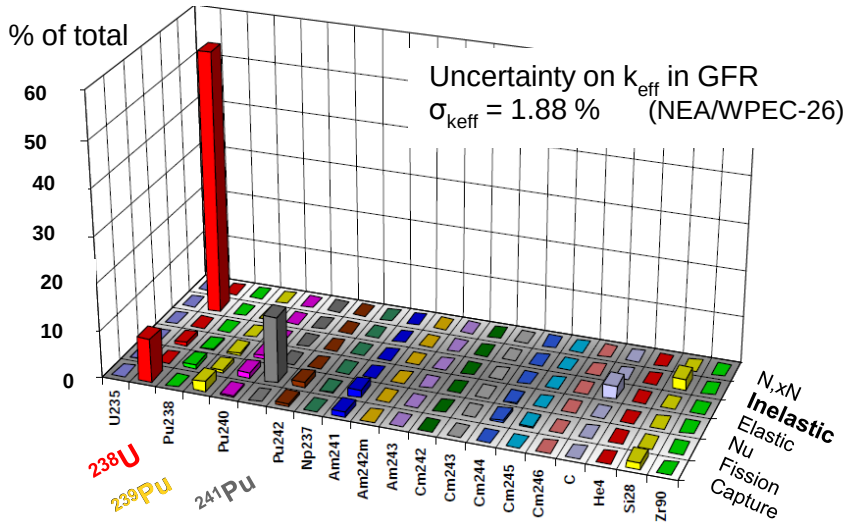


© Adapted from Wikimedia

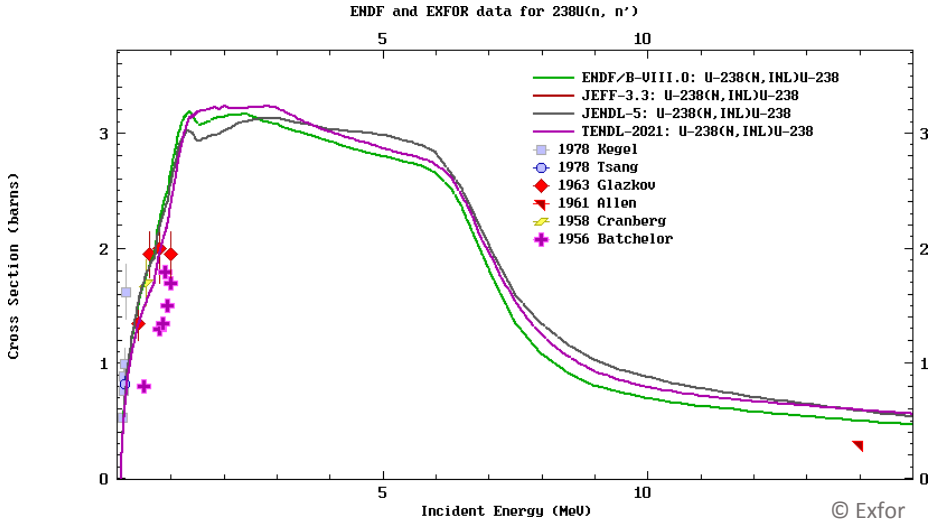
Inelastic Neutron scattering

The case of ^{238}U

- A nucleus of interest for current reactors, and future U/Pu fuels.
- Sensitivity studies show that current uncertainty on the computed k_{eff} in GFR (as in other Gen IV reactors) is dominated by the uncertainty on the ^{238}U(n, n') reaction cross section.



NEA, International Evaluation Co-operation, Volume 26 (2008)



© Exfor

- Indeed, current knowledge on ^{238}U(n, n') is still limited.
- Models and evaluations do their best given the lack of experimental data.
- ^{238}U(n, n') in the NEA's High Priority Request List.

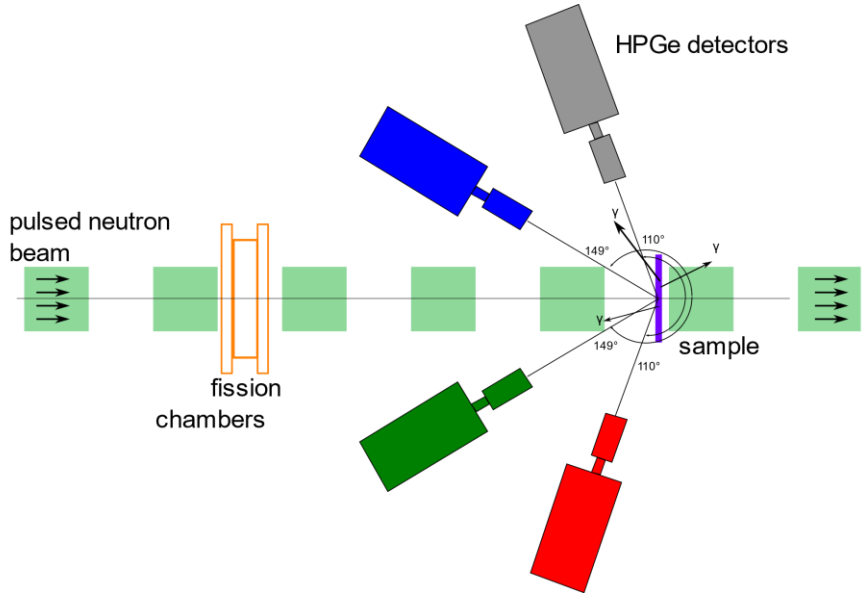
Request ID	18	
Target	**Reaction and process**	**Incident Energy**
92-U-238	(n,inl) SIG	65 keV-20 MeV
Field	**Subfield**	**Created date**
Fission	Fast Reactors EFR,SFR,ABTR...	28-MAR-08

••• Measurements via the exclusive channel (n, n' γ)

Measuring (n, n'γ) cross sections with Grapheme

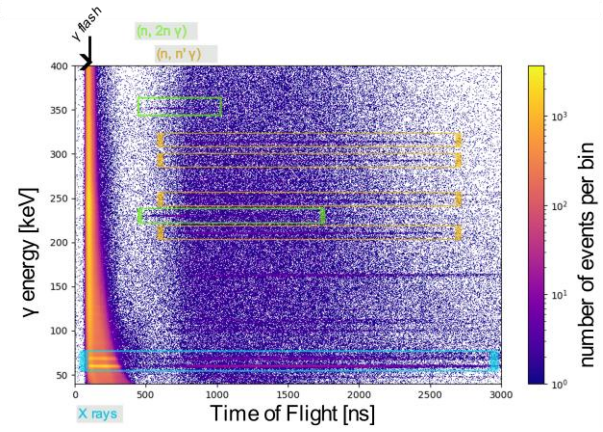
M. Kerveno et al,. EPJ Web of Conferences 239, 01023 (2020)

- At the pulsed neutron beam Gelina facility (JRC-Geel).
  - Continuous neutron energy distribution from keV to ~20 MeV
- Flight path: 30 m → Deducing the neutron energy by time of flight.
- Fission chambers to measure the incident neutron flux.
- 6 planar HPGe (including a 36-pixels one) with high efficiency and resolution for low energy γ rays (100-300 keV).
- Measurements done (or in progress) on ^{235}U, ^{232}Th, ^{nat,182,183,184,186}W, ^{238}U, ^{nat}Zr, ^{233}U, ^{57}Fe, ^{239}Pu



Identifying reaction channels and neutron energy

- Data recorded and displayed in γ energy vs. Time of Flight map.
- HPGe detectors allow a clear separation of γ rays
- Nuclear structure information is used to identify the transitions, signing unequivocally the reaction channel.
- Using the beam pulse signal and detection time, the incident neutron energy is determined by time-of-flight.



Measurements via the exclusive channel (n, n' γ)

Determination of cross sections

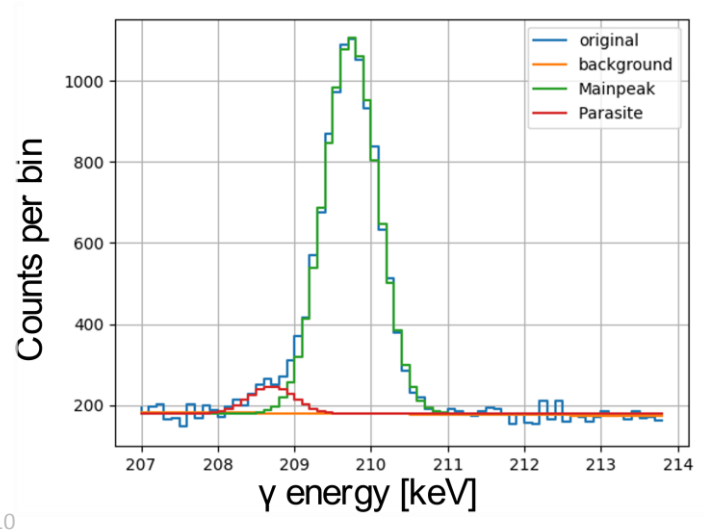
- Counting γ events at given time of flight.
- Angular cross section is obtained:

$$\frac{d\sigma}{d\Omega}(E_n,\gamma;\theta) = \frac{1}{4\pi}\,\frac{1}{N_{\text{target}}}\,\frac{N_\gamma(E_n;\theta)}{\varepsilon(E_\gamma)}\,\frac{1}{N_{\text{n}}(E_{\text{n}})}$$

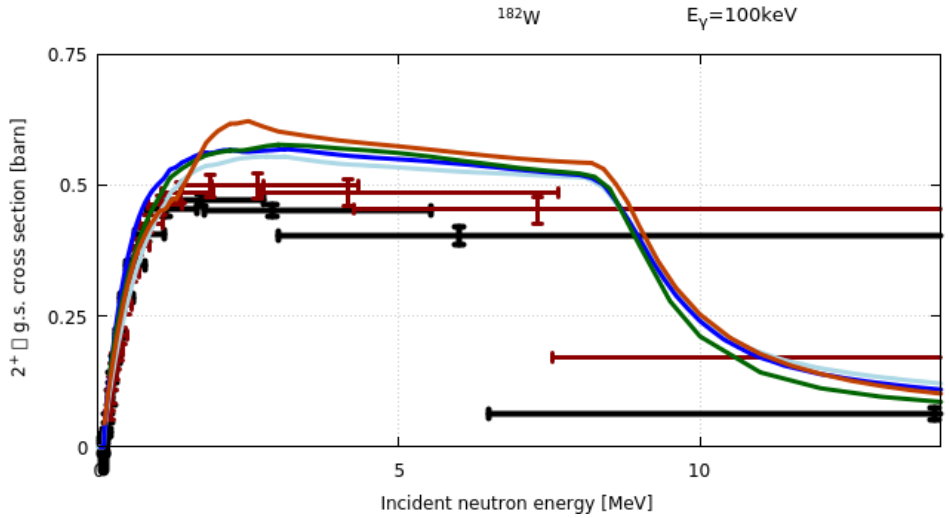
- HPGe detectors are placed at chosen angles (110° and 150°)
- Angle integrated cross section is obtained by linear combination of angular ones

$$\sigma(En,\gamma) = 4\pi\left(w_{110^\circ}\frac{d\sigma}{d\Omega}(E_n,\gamma;110^\circ) + w_{150^\circ}\frac{d\sigma}{d\Omega}(E_n,\gamma;150^\circ)\right)$$

C.R Brune, "Gaussian quadrature applied to experimental -ray yields". NIM A4 93 (2002) pp 106-110



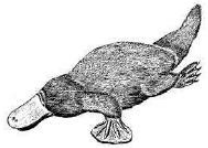
- Produces the angle integrated (n, n' γ) cross section as a function of incident neutron energy.



••• Measurements via the exclusive channel (n, n’ γ)

Deducing the total (n, n’) cross section from partial (n, n’ γ)

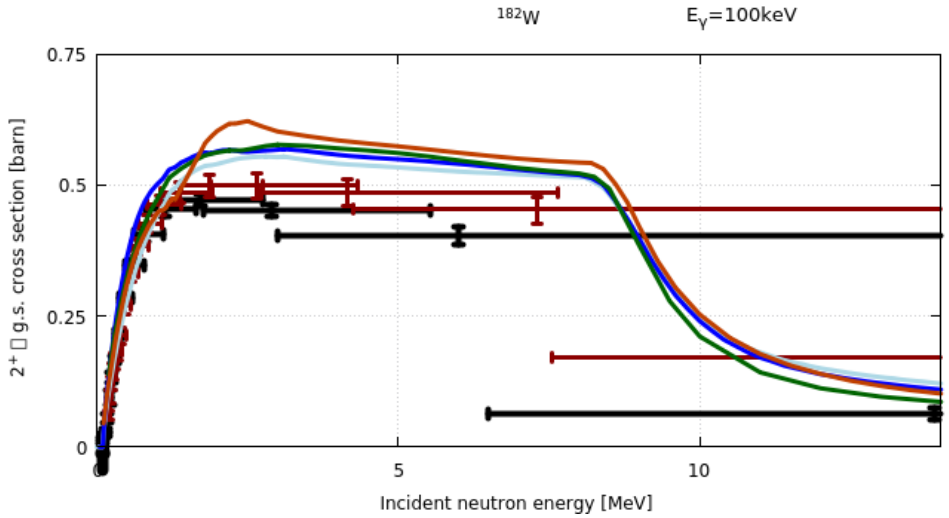
TALYS-1.8
A nuclear reaction program



© Talys Documentation

Reaction models and calculation codes.
(input: structure, masses, optical potential, level density, …)

Precise experimental (n, n’ γ) cross sections



Total (n, n’) cross section, computed with the models constrained by the experimental values.

This method relies on reaction models, calculation codes and structure information.
Precision and control of the uncertainties in the experimental data is very important.

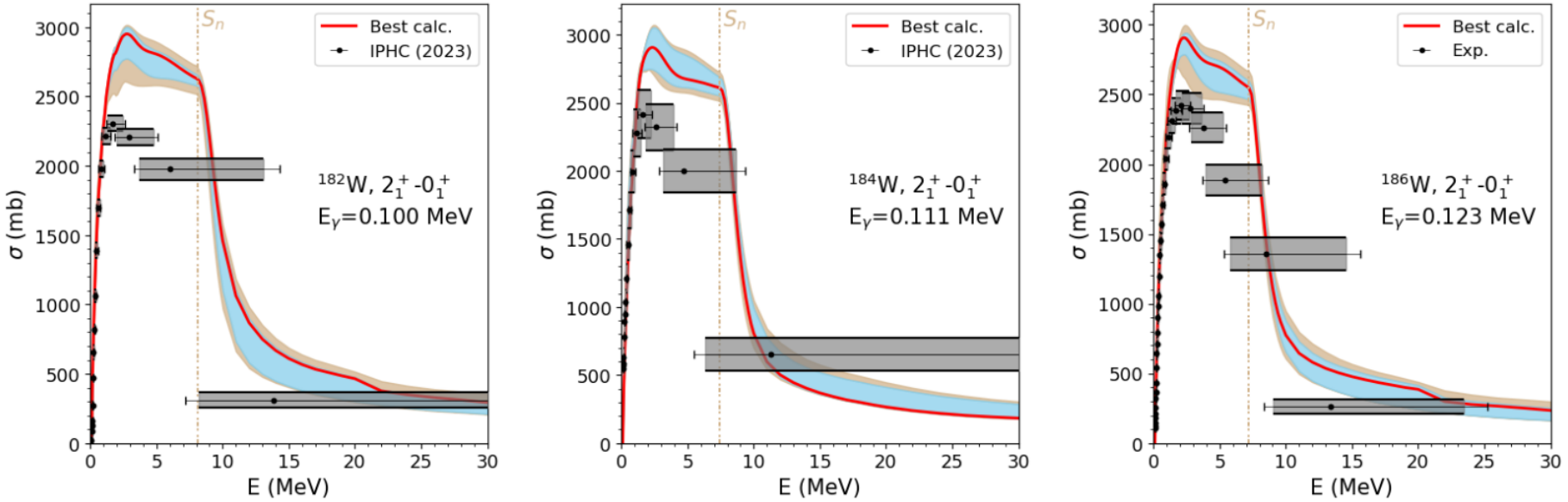
Looking at Tungsten

Interests

- Application: present in many alloys in reactors (fission and **fusion**), because of its mechanical and chemical properties.
- Theoreticians: Deformation of W isotopes similar to the one of actinides, no fission.
- Experimentalists: Not radioactive or toxic, no activation

Part of a large research program: data from $^{nat,182,183,184,186}W$

- Study of (n, n' γ) cross sections in $^{182,\ 184\ and\ 186}W$

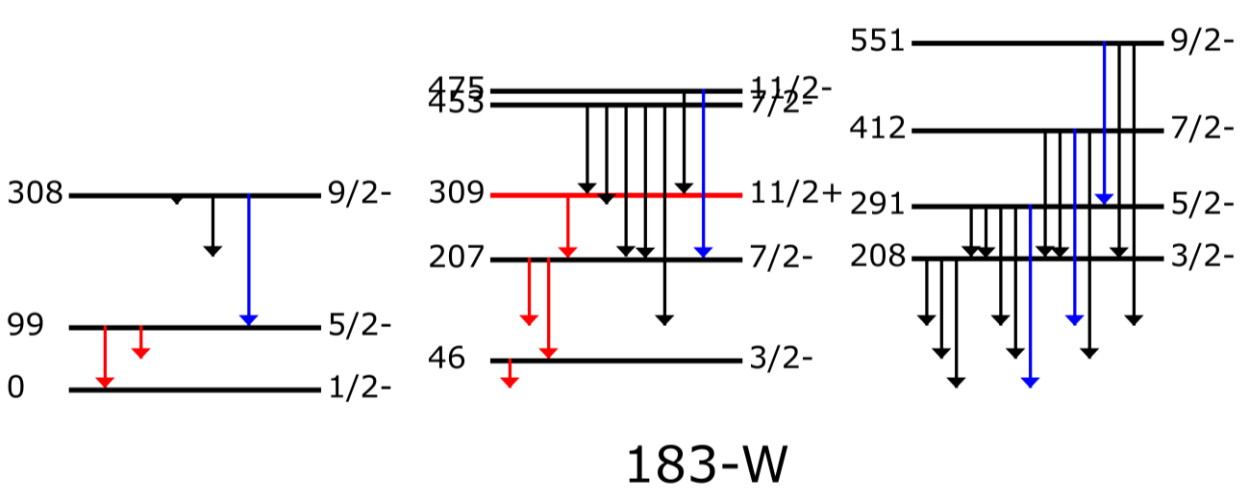


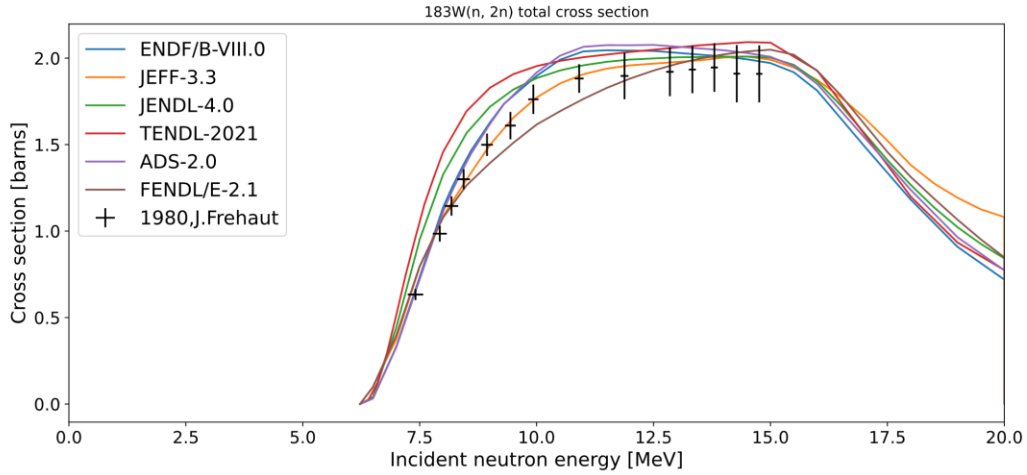
"Improving the accuracy of $^{182,184,186}W$(n, n') cross sections calculations" G. Henning, M. Dupuis, et al., In prep (2024)

Looking at Tungsten: ^{183}W

^{183}W

- Stable isotope, 14.3 % of natural W.
- Has one 5.2 seconds isomer at 309 keV
- Current knowledge of inelastic neutron scattering is sparse: mostly (n, 2n) (14 data points), a few (3 data points) partial (n, n') cross sections



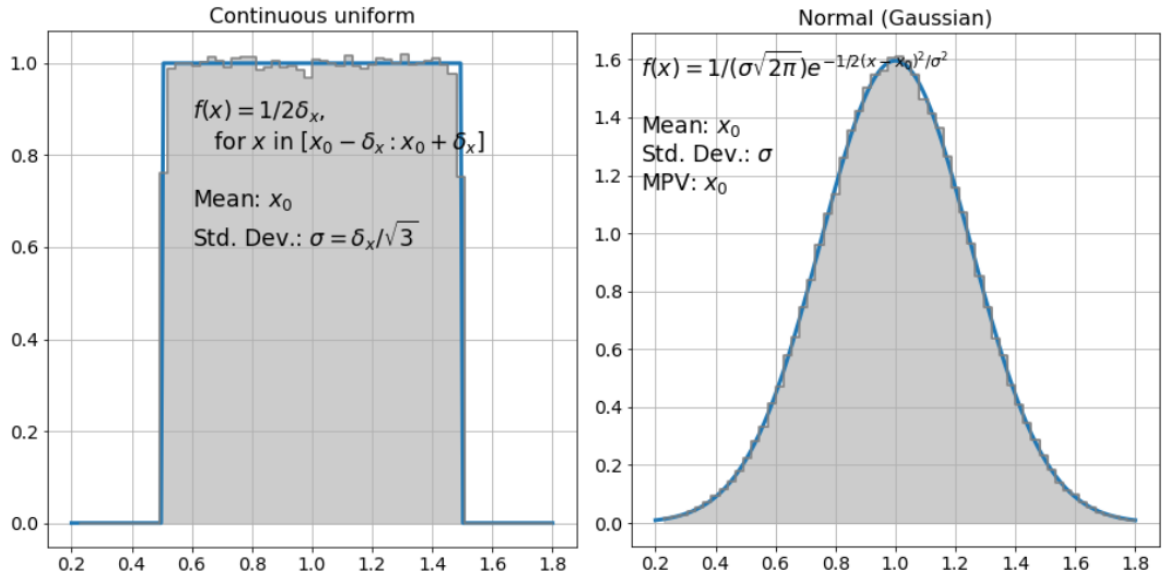


^{183}W(n, n' γ) and (n, 2n γ), along with already extracted $^{182\ \&\ 184}$W(n, n' γ) and upcoming ^{184}W(n, 2n γ) will make it possible to *thread* the study of inelastic scattering reaction from ^{184}W down to ^{182}W with overlapping channels.

Estimating uncertainties

What's "uncertainty" ?

- It expresses the level of confidence in the measurement results.
- Can be characterized by a range of values, a standard deviation, or a probability distribution.
- Formally, we consider our result x is one realization of a random variable X, defined by its probability density function, with a mean value $\bar{x}$ and a standard deviation σ_x.



Combination of uncertainties

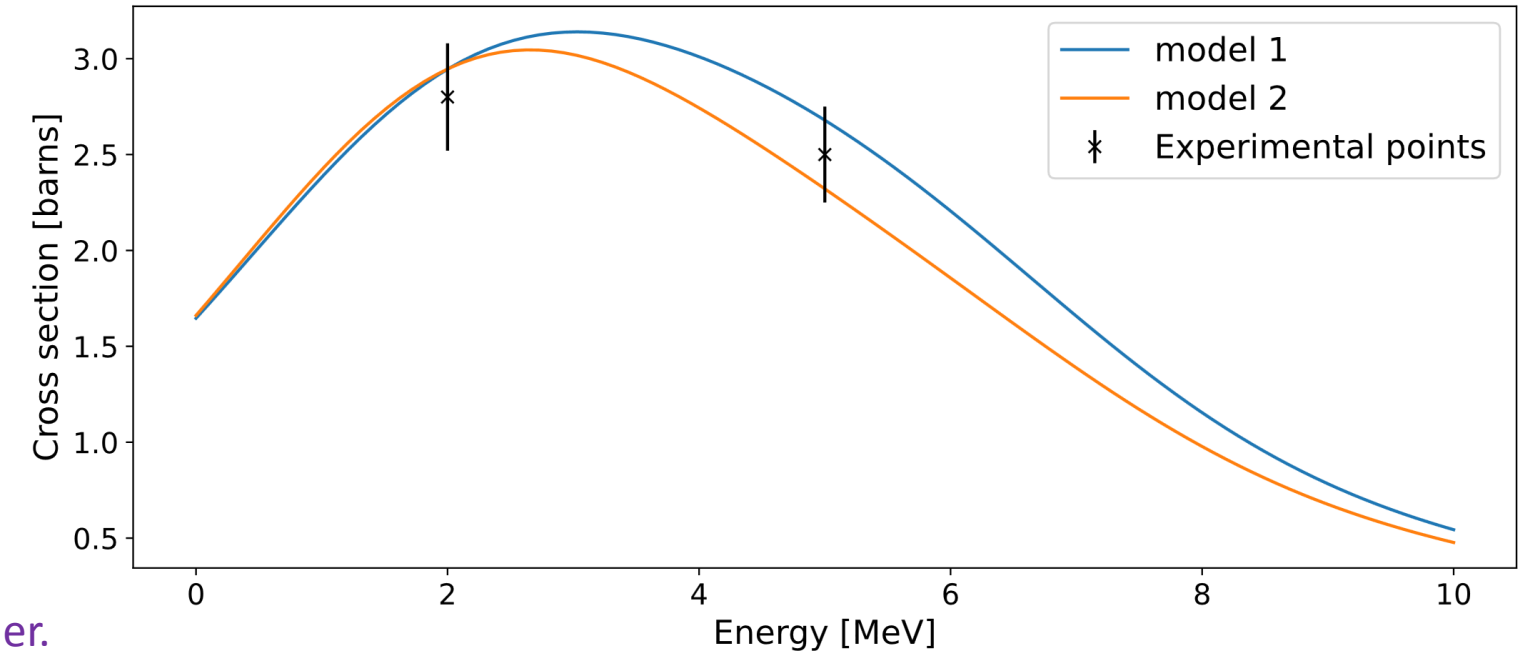
- The combination of uncertainties is usually done with the *square summation* formula (from perturbative theory).
  If $a = f(x, y)$, with x, y two independent variables :

$$u_a{}^2 = \left(\frac{\partial f}{\partial x}(\bar{x}, \bar{y})\right)^2 \times u_x{}^2 + \left(\frac{\partial f}{\partial y}(\bar{x}, \bar{y})\right)^2 \times u_y{}^2$$

- When the variables are not independent, an extra term (possibly <0) appears: $+2 \times \frac{\partial f}{\partial x}(\bar{x}, \bar{y}) \times \frac{\partial f}{\partial y}(\bar{x}, \bar{y}) \times \langle (x - \bar{x})(y - \bar{y}) \rangle$
- $cov(x, y) = \langle (x - \bar{x})(y - \bar{y}) \rangle$ is the covariance between x and y.
- $corr(x, y) = cov(x, y)/(u_x u_y)$

Why covariance matters

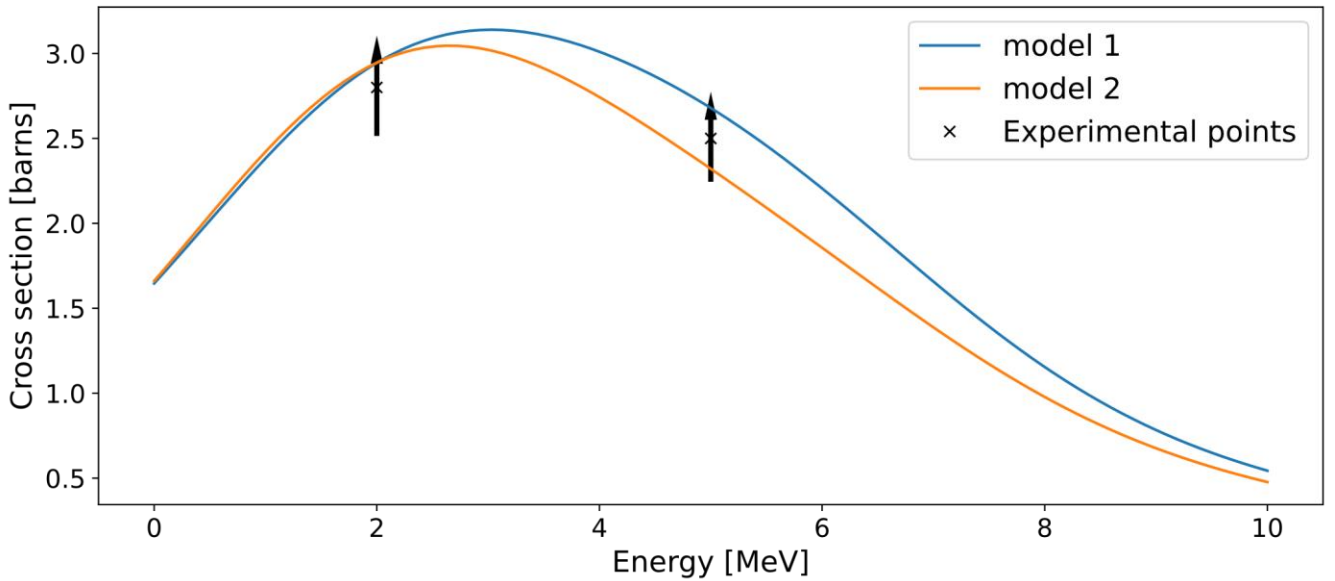
- Let's take an hypothetical situation with two experimental points, compared to two model calculations.
- The compatibility of the models with experimental data can be expressed with a χ^2 calculations
- Here $\chi^2_{\text{model 1}}=0.782$, $\chi^2_{\text{model 2}}= 0.784$



No way to favor one model or the other.

Why covariance matters

- Adding the information of covariance, it “links” values together,
- Correlation comes typically from **common parameters** in the data analysis.
- Can be illustrated as arrows on the uncertainty bars.
- Generalized χ^2 with covariance $((X - \bar{X})^T Cov^{-1}(X - \bar{X}))$, is now (with $corr_{1,2}$=0.95) 0.0002 for model 1 and 1.5704 for model 2



No new measurement, no change in the experimental values, or their uncertainty.
Just by taking into account the correlation between points, one model can be hugely favored compared to another.

It is very important for evaluation.

High Priority request	
Target uncertainty	**Covariance**
See details	Y

As covariance is driven by analysis parameters, as much information as possible should be given with the experimental results: values, uncertainties, covariance, and how they were obtained.

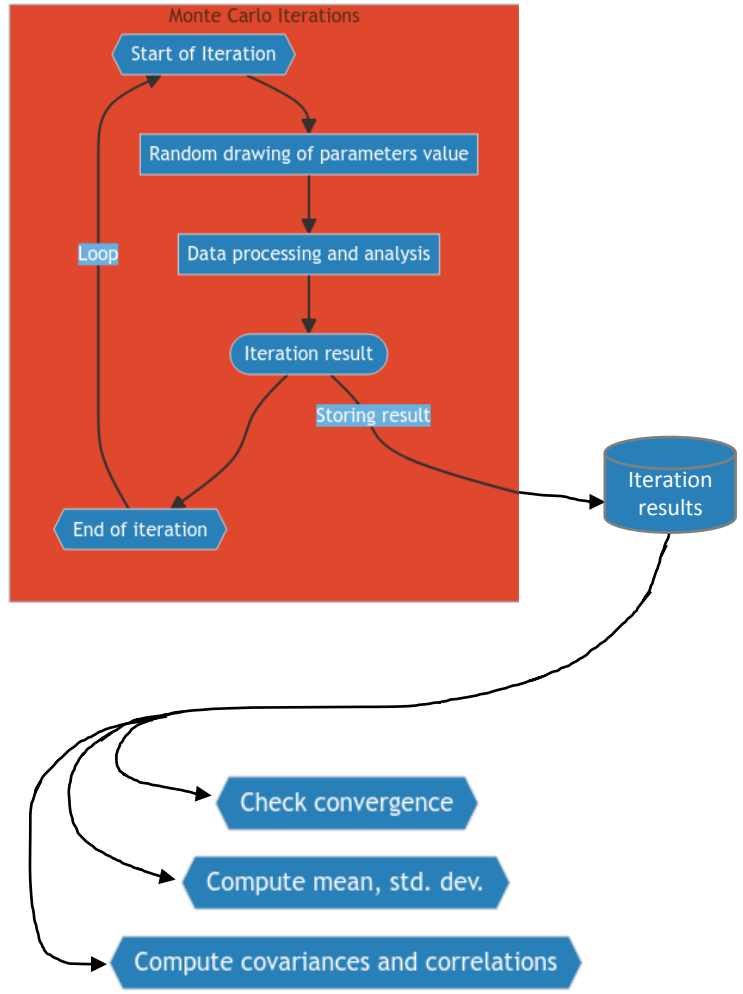
••• Using Random Sampling (a.k.a. Monte Carlo)

Why use random sampling?

- Uncertainty and covariance can be computed using analytical formulas.
- When there are many parameters, the mathematical derivation can become tedious. And a source of mistake when transcribing it into a computer program.
- Others strategies are possible… among which, Random Sampling (a.k.a. Monte Carlo)

What is random sampling?

- Parameters are sampled randomly along their probability distribution
- Analysis is performed
- Results are stored onto a *stack*
- After N iterations,
  - Check for convergence, i.e. *did I gather enough samples to have a meaningful result ?*
  - The stack of results is processed to computed mean values, standard deviation, correlation, … using statistical formulas (or defer to well established libraries).



••• High quality nuclear data with Open Science methods

For evaluation, as much information as possible should be given with the experimental results: values, uncertainties, covariance, and a detailed description of how they were obtained.

Open Science is the good framework for such extensive documentation

Open science is built on three pillars :

- Open Access (to publications)
- Open Data
- Open Source

Open Science framework is also a great tool to ensure the continuity of research
(i.e. resilience to changes in team composition, and/or computer crash)

••• High quality nuclear data with Open Science methods

Open Access

- Not just a question of price, but also rights to reuse (tables, figures, …) and distribute article content.
- Different models: Who pays? Who distributes?
- ANR or EU funded projects require Open Access publications
- CNRS researchers evaluation is based on articles made available in open access on the HAL platform.
- In France, since the 2016 “Loi pour une République Numérique”, all publicly funded research articles can be published as *postprint* in open access after an embargo period of 6 months, regardless of the (maybe commercial model of the) editor of the article.

What to publish?

- Peer-reviewed articles, for sure
- But Open Access platforms like HAL also accept Posters, Images, Working papers, …

Why?

- Meet funding/evaluations requirements
- Establish *ownership,* precedence
- Ensure research continuity

High quality nuclear data with Open Science methods

Open Data: Making research data freely available for others to use and reuse

Be FAIR

- Findable: by depositing the data in a recognized repository that will assign it a DOI, and give a correct and full description of the data with extensive *metadata*
- Accessible: via the repository. In France, Recherche Data Gouv, in Europe Zenodo
- Interoperable: Metadata should include full description of the data format or required software to read it.
- Reusable: The data is published with a specific "License" that states what can be done with the data, and under which conditions it can be used.

  (Typically, CC-BY **allows** the copy and redistribution, as well as *building upon* the material, while **requiring** citation of the original work.)

Metadata

- "all the information that is necessary to identify and make sense of Open Data files"
- Include Authors, funding acknowledgment, data description, license, …
- Included in a specific computer readable format for efficient indexing, to make the data Findable.

Data Management Plan

A tool to prepare the data life cycle (collecting, analysis, archiving, publication) in advance.

••• High quality nuclear data with Open Science methods

Open Source: Sharing (analysis) code

Publishing the analysis code

- The analysis code is as much part of the scientific process leading to the results than any theory, experimental setup, parameters, …
- Publishing an analysis code in Open Source is a way to ensure code quality, allow others to reuse (part of) it (so that they spend less time and energy re-creating already existing software parts), and great for research continuity.
- More importantly, it *shows* how the scientific results is obtained and builds trust in the results.

Writing Open Source Code that others will understand

- Use versioning system (such as `git`)
- Include documentation, license, …
- Organize the code directories
- Ensure that the required libraries/dependencies needed to run the software can be found/installed (use virtual environments or containers).

Publication

On `git` repositories (`gitlab.in2p3.fr`), general open science repository (HAL, Zenodo), or/and Package managers (`Pypi`, `cargo`, `npm`).

••• High quality nuclear data with Open Science methods

For evaluation, as much information as possible should be given with the experimental results: values, uncertainties, covariance, and a detailed description of how they were obtained.

Publishing data and analysis code along side the scientific article, is the best way to reach the level of detail in the documentation on how the data is analyzed, and the uncertainties produced, that is required for accurate and meaningful evaluation of experimental values.

As a bonus, it is also a key element to maintain the continuity of research in a team.

Team work organization should also be adapted to Open Science goals

- Use of document/data/code sharing tools (`git`, …)
- Task list, with fine division in subtasks.
- Regular feedback on results and how to improve the work methods.
- Emphasis on simplicity, being explicit, practical → documents will be read, codes will be run, … in the distant future.

••• Implementing a full Monte-Carlo analysis code

Why a new code?

- Previous codes were not producing covariance, correlation matrices
- & were developed without long term thinking, and became too complex to maintained

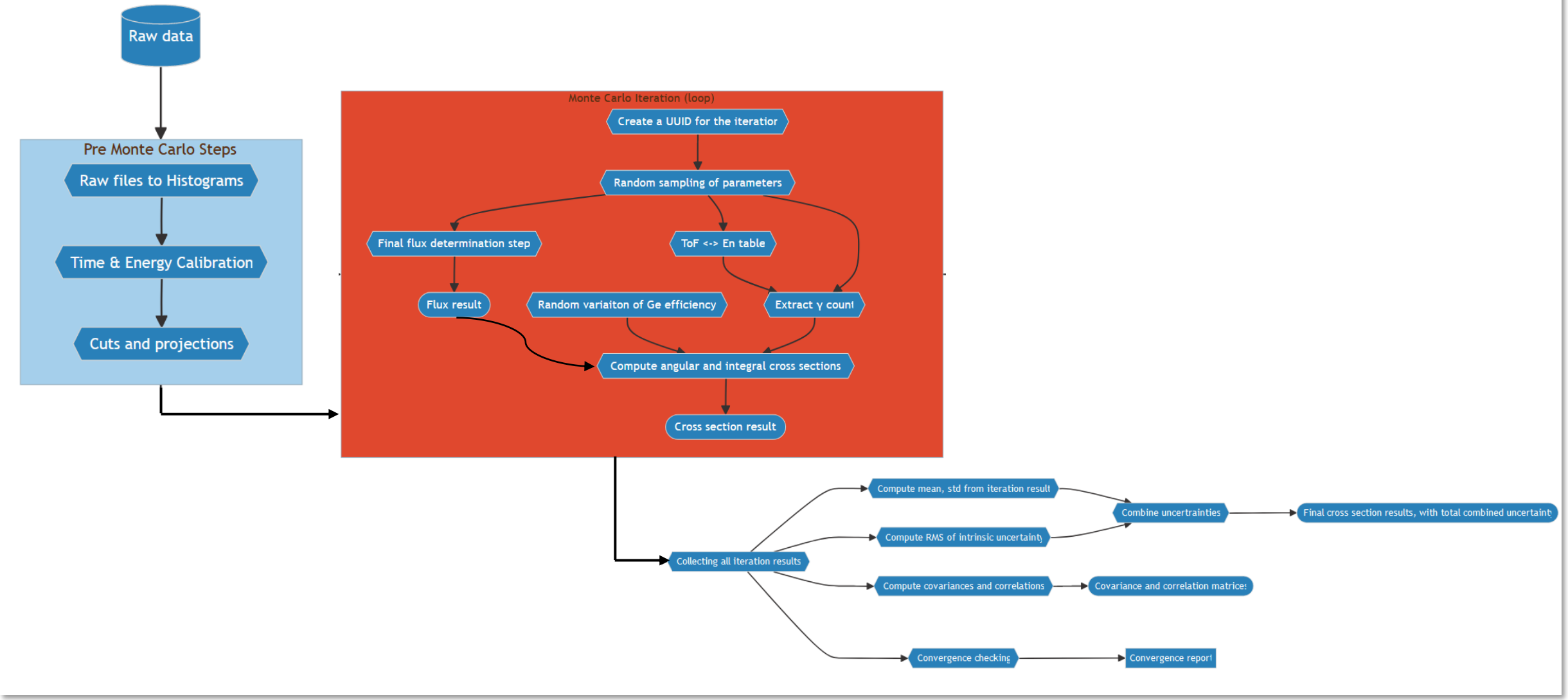
- New code will
  - Produce uncertainty and covariance, using random sampling
  - Built as a *pipeline* and keep all intermediate results for debugging
  - Be portable between platforms
  - Be flexible, adaptable to future experiments (change in number of detectors, raw files format, …)
  - Be published in open source

How is it built?

- Coded in `Python` (because of its high portability, and ease of creating new scripts)
- Uses `Yaml` for input, outputs are all ascii files (high readability by other programs)
- Uses `Numpy` for mathematical operations (a validated, trusted, and open source library)
- Uses `doit` for job management (à la `Makefile`)
- Runs on IN2P3's Computing Center, but should be able to run on any platform with minimal adaptation.

••• Implementing a full Monte-Carlo analysis code

Analysis flowchart

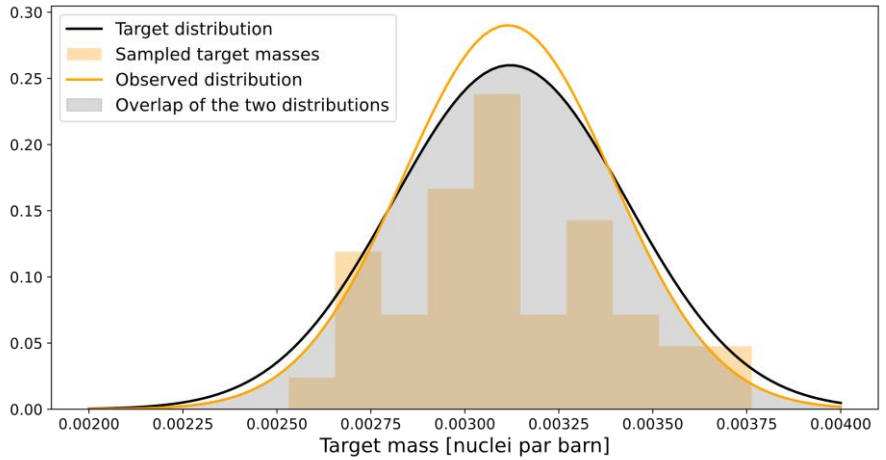


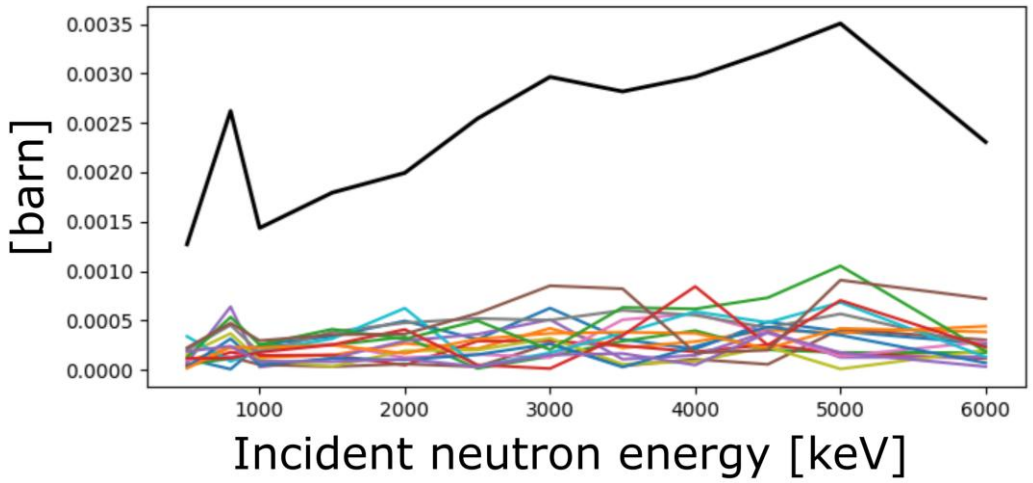
••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

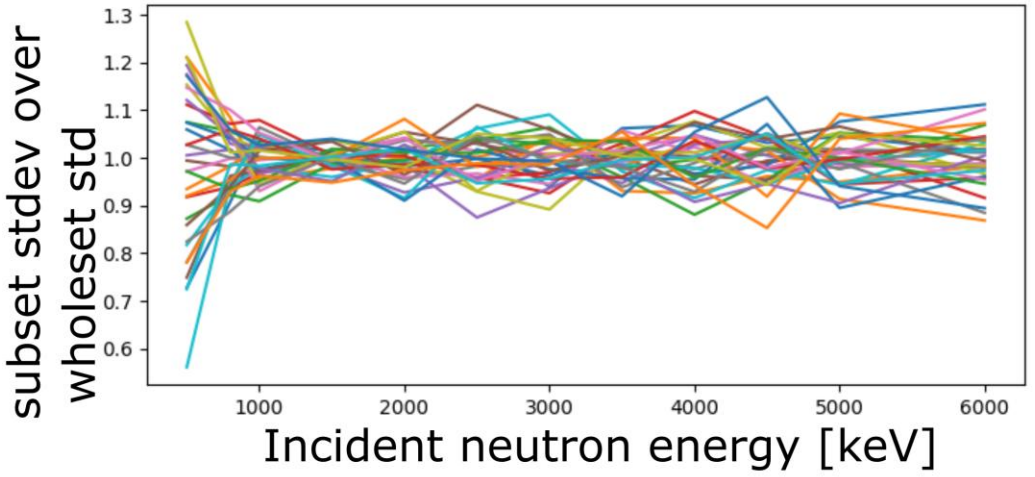
First, checking for convergence

- *"Did I gather enough samples to have a meaningful result ?"*
- For input sampled parameters: does the sampled value correctly reflect the intended distribution ?

- For outputs, by dividing the results in subsets, we verify that any additional iteration is unlikely to change the computed mean values and standard deviation.



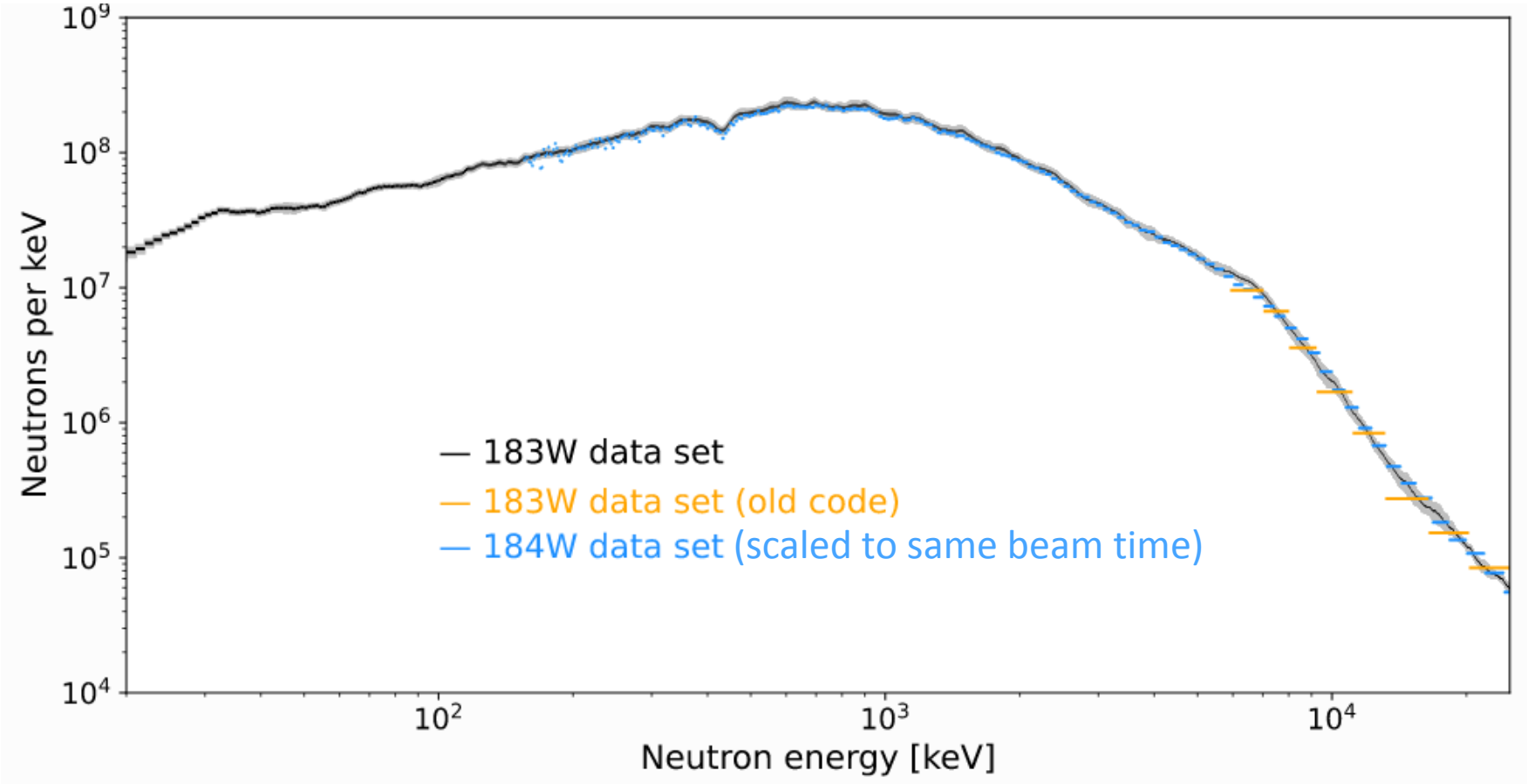




••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

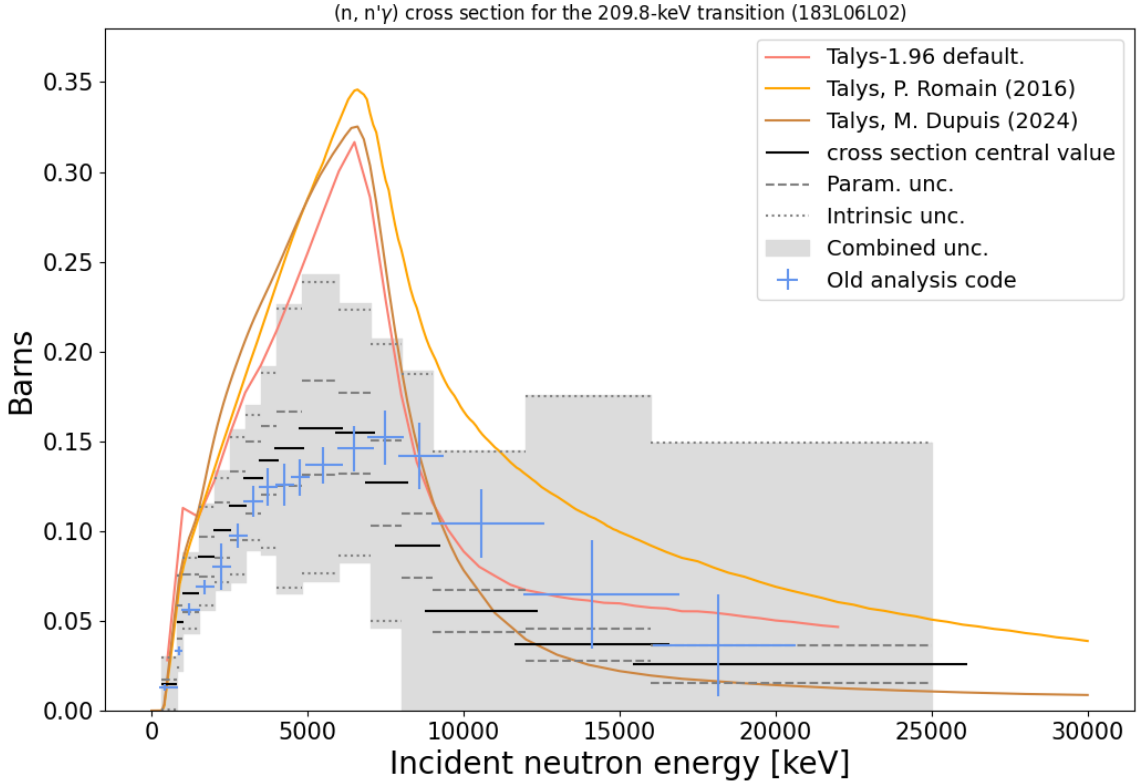
Checking result for the neutron flux

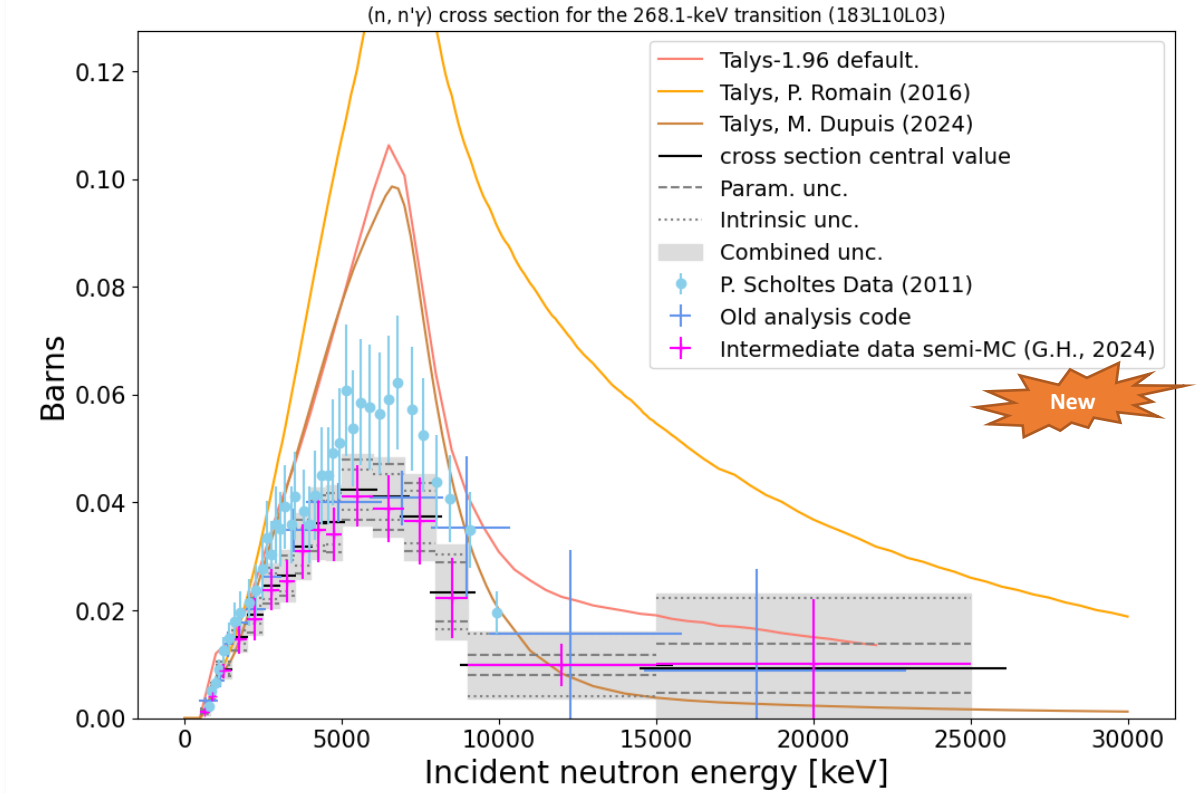
Compared to results from different data set, and other analysis code.



••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

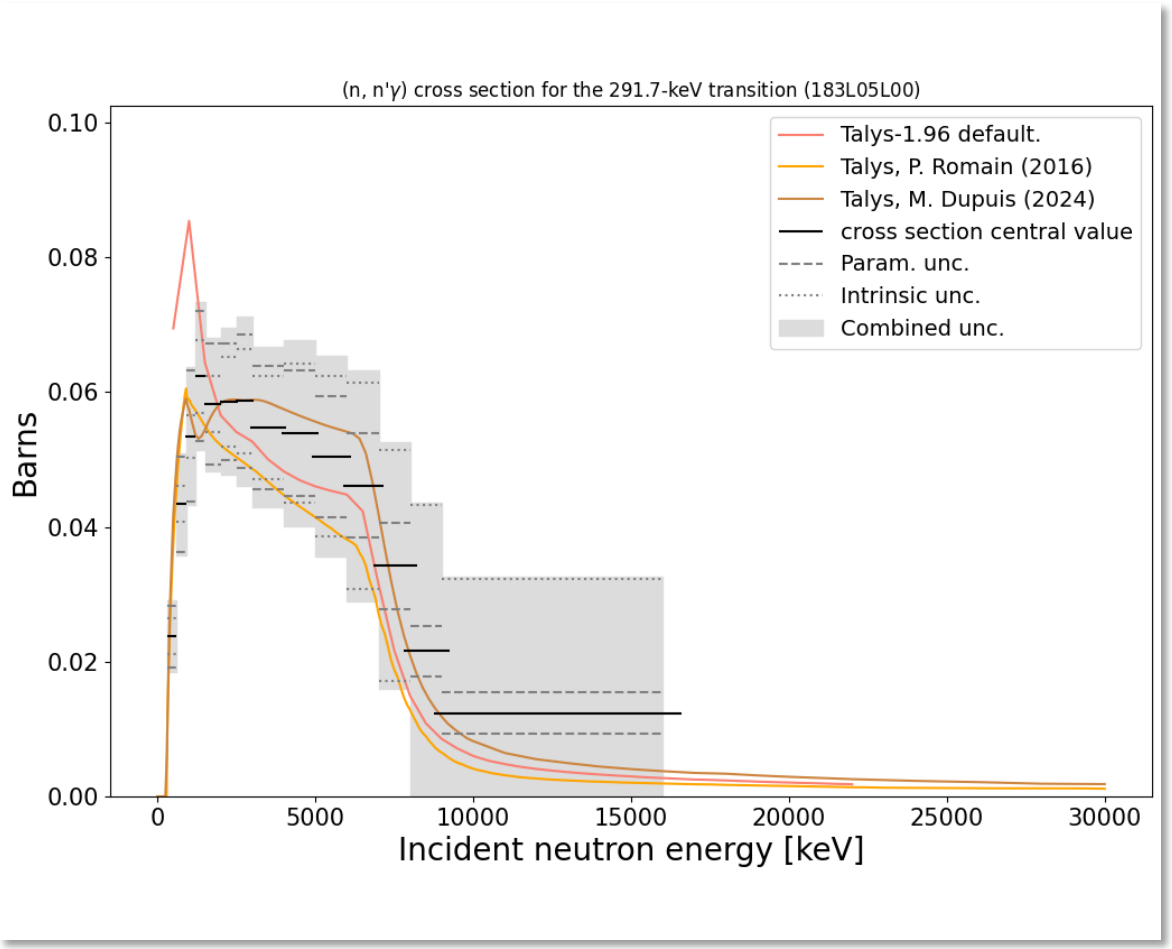
Compared to Talys calculations, and, when available, to results from the same data set obtained from prelimary analysis (P. Scholtes) and with the previous analysis code (?unsure about the corrections applied).

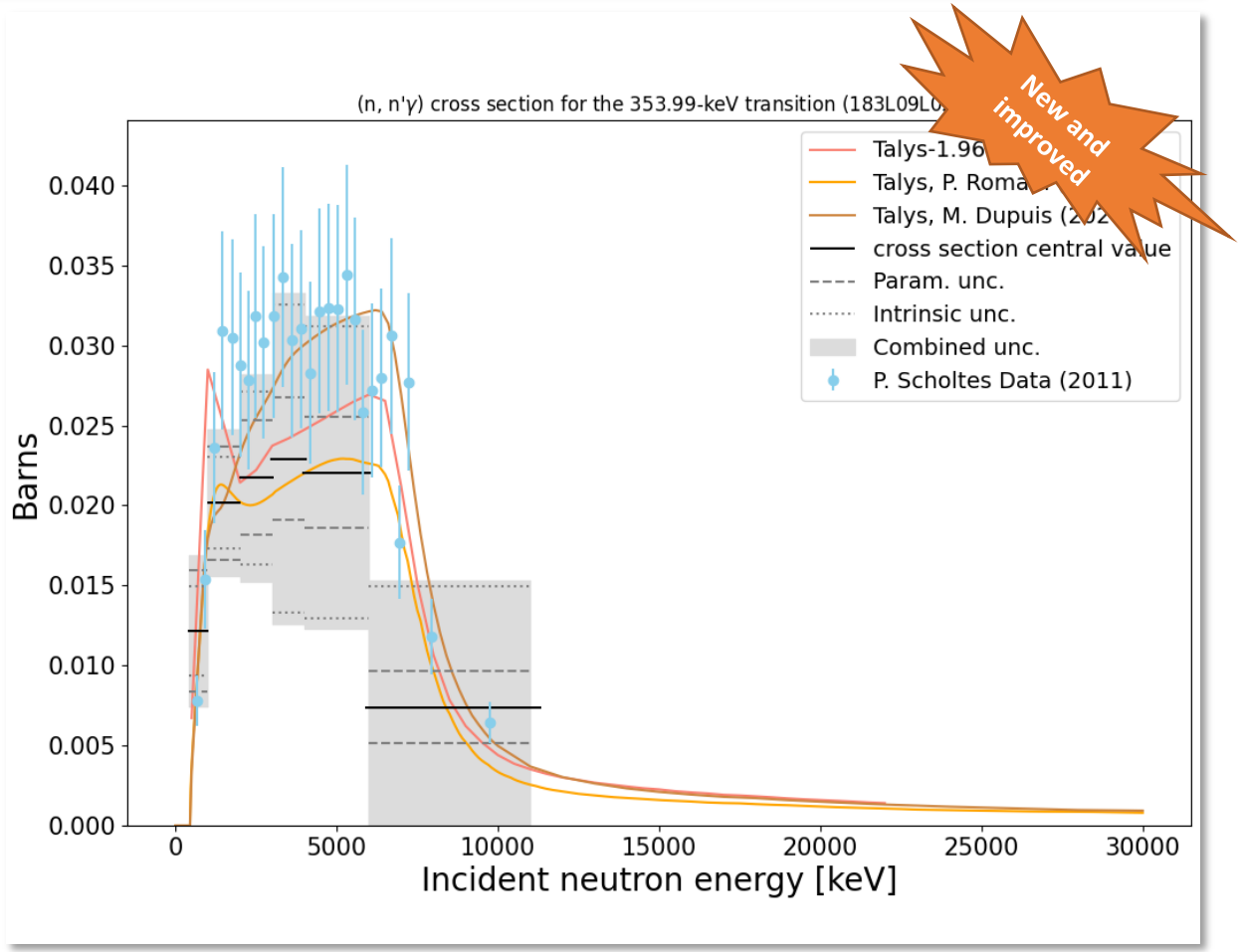




••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

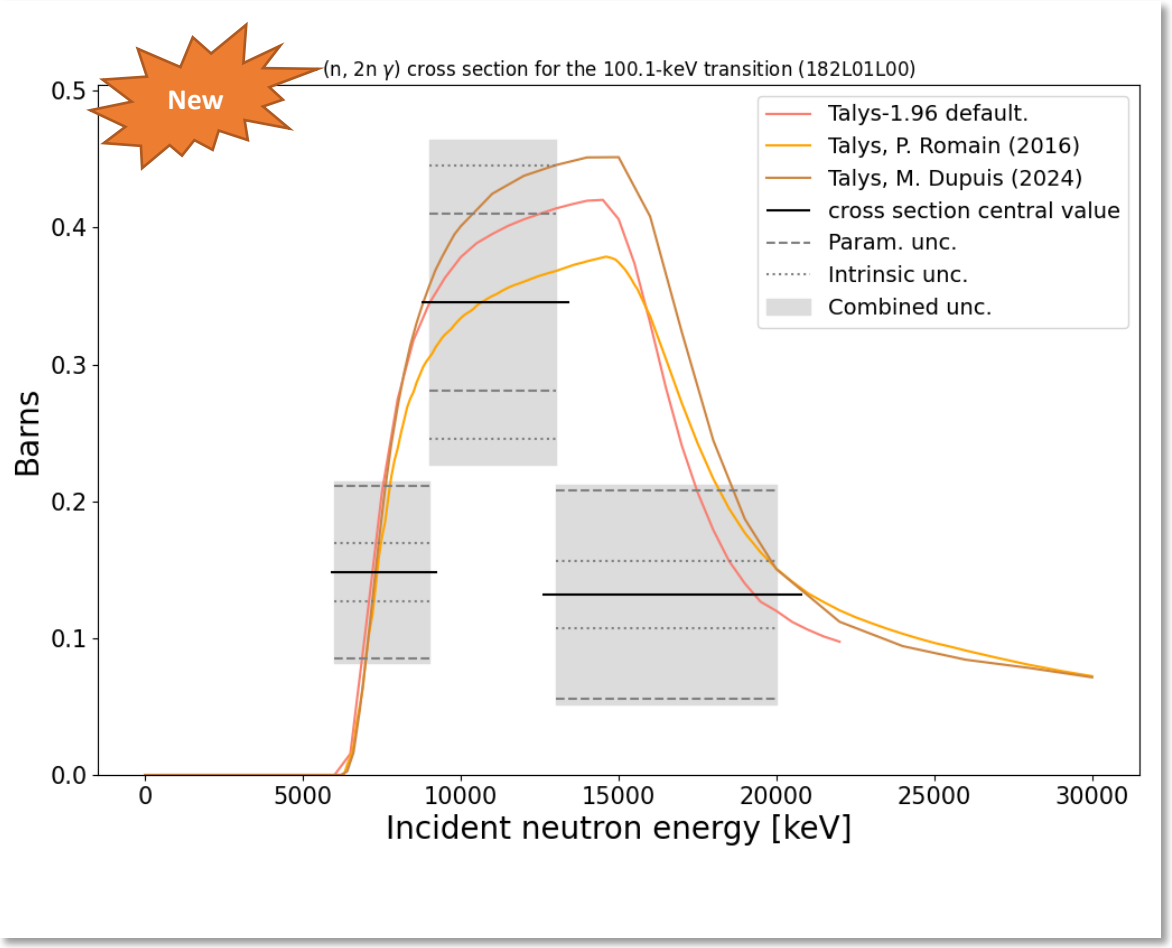
Compared to Talys calculations, and, when available, to results from the same data set obtained from prelimary analysis (P. Scholtes) and with the previous analysis code (?unsure about the corrections applied).

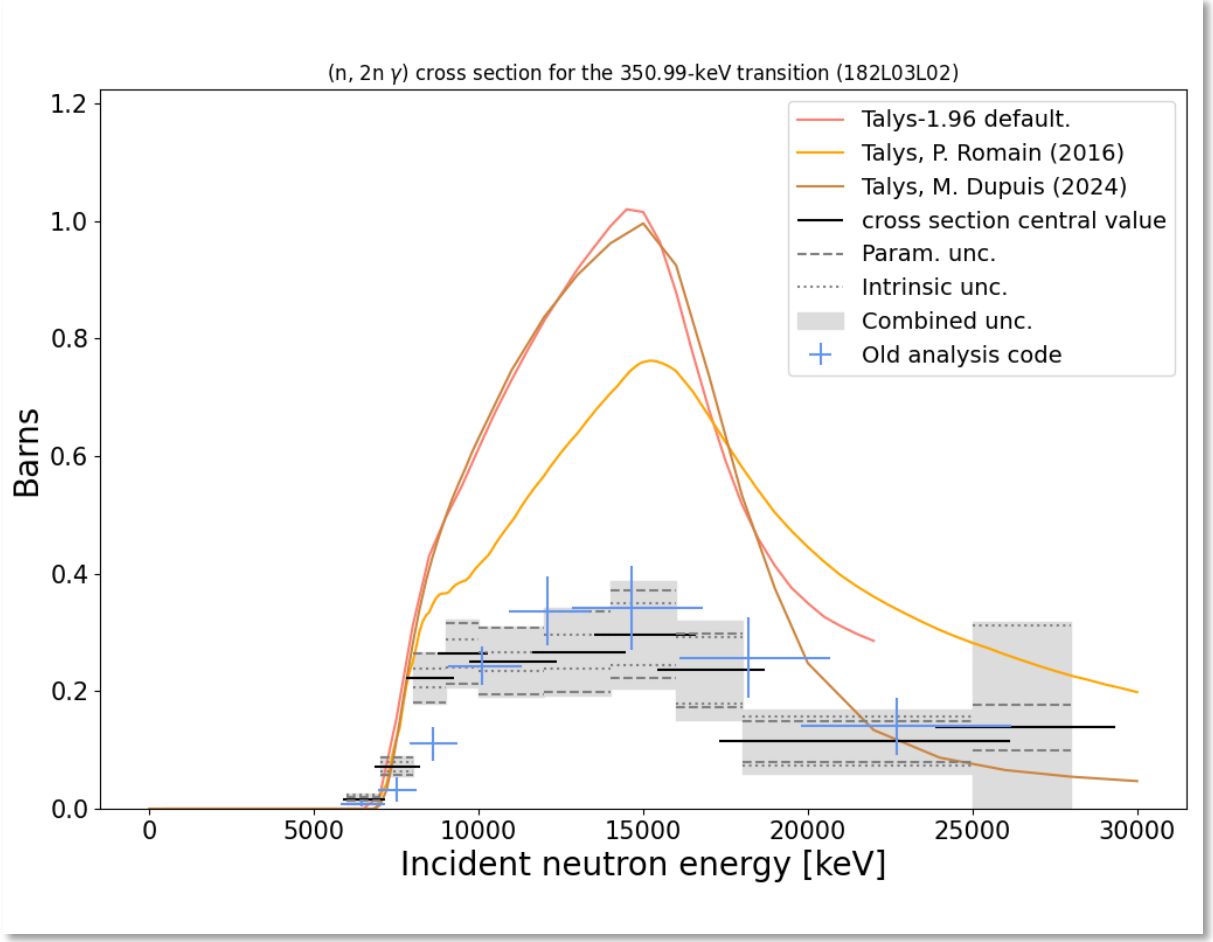




••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

Compared to Talys calculations, and, when available, to results from the same data set obtained from prelimary analysis (P. Scholtes) and with the previous analysis code (?unsure about the corrections applied).

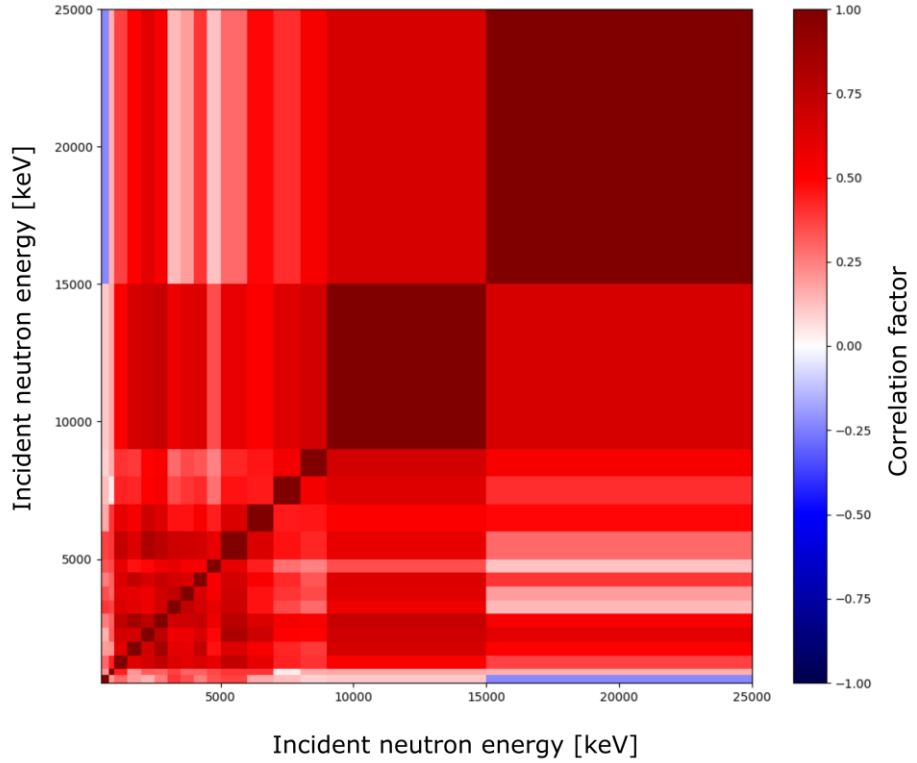


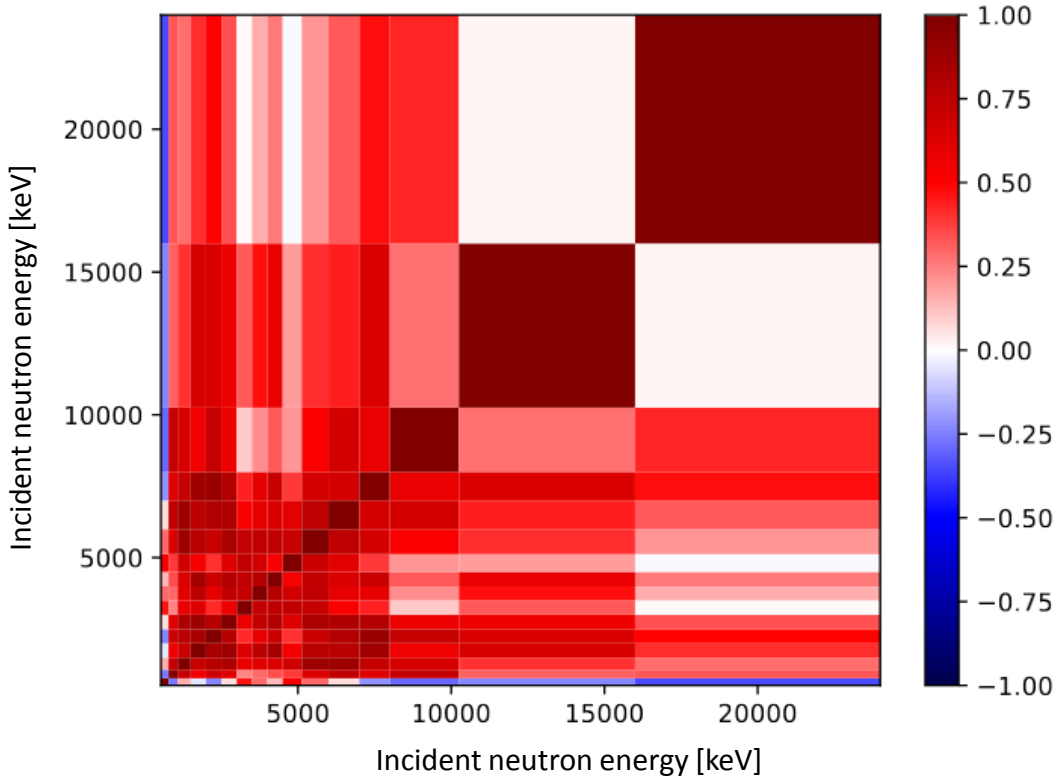


••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

Correlation matrices

Here for 268 keV γ — full MC (left) compared to the results from semi MC calculations using intermediate data (right).





The differences in correlation matrices reflect the differences in data processing → importance of an open code.

Opens de possibility of using the semi MC code (the 1st code we developed in a sustainable way in collab. with F. Claeys, also used by N. Dari-Bako) for analysis.

G. Henning, F. Claeys, et al. EPJ Web of Conf., 294 (2024) 05002

••• Full Monte-Carlo ^{183}W(n, n' γ) and (n, 2n γ) results

Discussion

- Confidence in results, thanks to compatibility with calculations, and other codes (given their uncertainty)
- Choice of fission chambers still to be considered.
- Some improvements and fixes already implemented (compared to the version presented in the manuscript):
  - Fit: fixed initial guess for background component.
  - In code: `context` module to handle analysis context information
  - House keeping in random sampling tasks

- Additional improvements coming up:
  - More improvement on Fit (aim: reduce intrinsic uncertainty)
  - Speed up Monte Carlo loop

- Plan: to use the code systematically with other data sets, including with different data source (Faster acquisition), number of detectors (Grapheme post 2015), at other facilities (NFS), and using other analysis methods (semi MC).

New code

- Created a new, full MC analysis code from scratch
- Provides uncertainties, covariance and correlation matrices.
- Designed in Open source, with extended documentation.
- Saves all intermediate steps in `ascii` format → great for debugging

- Will benefit from continuous improvement, by design.
- Not experiment/dataset specific: will be broadly use in the future.

New code, **New tool**, New method

New Tool

- Flexible analysis framework: can be adapted to other configurations (DAQ, number of detectors, …)
- Designed for long term use, in future experiments, datasets.
  - This will help maintain continuity in our results production.

- Can evolve with experiments, detectors, methods, … thanks to the pipeline architecture

- Will speed up data analysis, with a high degree of reproducibility, streamlined data analysis pipeline.
  - Analysis *life-cycle* will be adapted (data publication, …)

New code, New tool, **New method**

New method

Based on this experience, and *armed* with this new tool, we can adapt our work methods, with the aim of increasing our (small) team resilience, allow us to keep producing high quality physics results, maintain our level of expertise and skills.

Ensuring research continuity

- It has been an issue in the past (finding/understanding codes, or files left by students, after they moved to other position).
- We are already pushing for an increased use of shared documents, `git`, …

- Code quality & repository organization is a broader issue, regularly discussed between colleagues.
- There is an identified need for training on that topic, as early as M2 level and beyond (PhD students, postdocs, …)

New code, New tool, **New method**

Open Access Publication

- In response to funding or institutional requirements

- Not just for articles:
  - Data sets and associated data paper
  - Working papers for low impact or early stage research that does not (yet) warrant a full *article.* Working paper should be written at least for internal use (continuity), but it is even better to share it with the community.

- ☹ Open datasets, working papers , … are not included in researcher's performance evaluations in CNRS.

Open Analysis code

- The analysis code is an integral part of the experiment.
- *Missing link* between a well tested and described detector, and extensively discussed results.
- Important for the production of nuclear data and taking them into account properly for evaluation, but more broadly, it builds confidence in the results, and increases reproducibility.

Perspectives

^{182,183,184}W (n, n' γ) and (n, 2n γ) cross sections

- (n, n' γ) results on even-even isotopes will be published soon (article + Open Data)
- ^{183}W(n, n' γ) & (n, 2n γ) cross sections published in Open Data by the end of 2024, along with an article focused on data.
- ^{184}W(n, 2n γ) cross section: value extraction soon. Open data publication around end of 2024 / early 2025, with data paper.

- Will provide a significant cross isotope dataset for theoretical interpretation.

Using the analysis code on ^{239}Pu data

- After adaptation to different DAQ, number of detectors, compared to W dataset.

- Once the pipeline is in place, additional data (recording is ongoing) is processed *on the fly.*

^{238}U(n, 2n γ) and (n, 2n) cross section at NFS

- Using the same analysis framework
- Ganil data policy will apply
- A Data management plan has been written

Perspectives

Many low-key, exploratory subjects ready to go (?)

- Grapheme fission chambers differences.
- Looking into Exfor entry 41245.022. ?

- Using machine learning and large models for data interpretation (M2 computer project, 2023). ? ?
- Al neutron induced activation to estimate neutron flux (or background) in our experimental areas. ?
- Background characterization in the Grapheme experimental hall (M2 internship subject 2022). ?

- Geant4 / MCNP6 comparisons.

- Extracting compound nucleus spin distribution from experimental cross sections.
- Developing a portable DAQ system (M1 internship subject in 2021).

These projects can be kickstarted with internship, computer projects, … (leveraging continuity tools is therefore important)

Take Away Message

- There is a need for improved nuclear data evaluations for the development of nuclear applications, requiring new measurements and better theoretical descriptions by models.
- The group DNR at IPHC focuses its work on (n, xn γ) reactions, with the measurements of cross sections.
- I presented the first results for ^{183}W were obtained with a new, Full Monte Carlo code.
- The detailed explanation of how the data is produced is important for meaningful evaluation and the principles of Open Science (Open source code, Open data) are the best way to provide the transparency needed for proper exploitation of our results.
- This Open Science approach, as well as methods to help conform to it, will be our guiding principle in the future.

••• You get everything !

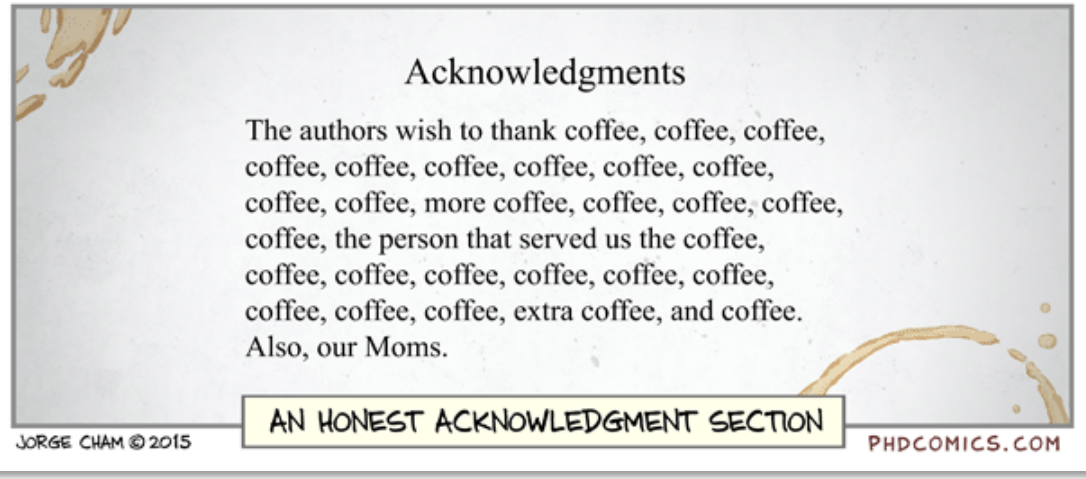
Embracing the principles of Open Science, all the work and files related to this HDR will be distributed openly.

- The manuscript (PDF and HTML pages),
- The source code of the manuscript (`.rst`) including the images and scripts to replot the graphs and flowcharts,
- The scripts to compile the manuscript sources in a PDF or HTML pages,
- Some scripts related to the manuscript (to draw level schemes, generate metadata),
- The data,
- The analysis code,
- This presentation (PDF and PPT files),
- The presentation video.

What else do you want?

Thank you

Maëlle and Philippe,
Members of the jury,
Past and present colleagues and collaborators at IPHC, with special thanks to the technical and administrative staff,
CSDDD members,
Colleagues at Université de Strasbourg, in particular the Open Science community,
Collaborators in the Eedin collaboration, at IAEA, Los Alamos,
Collaborators at JRC-Geel, IFIN, CEA, Ganil,
Collaborators in the Nacre project, GDR Scinee,
All the others,



••• Questions