AI data bases and material discovery

E. Farhi, *Data Reduction and Analysis Group - SOLEIL*

DIADEM School - Aug 2025

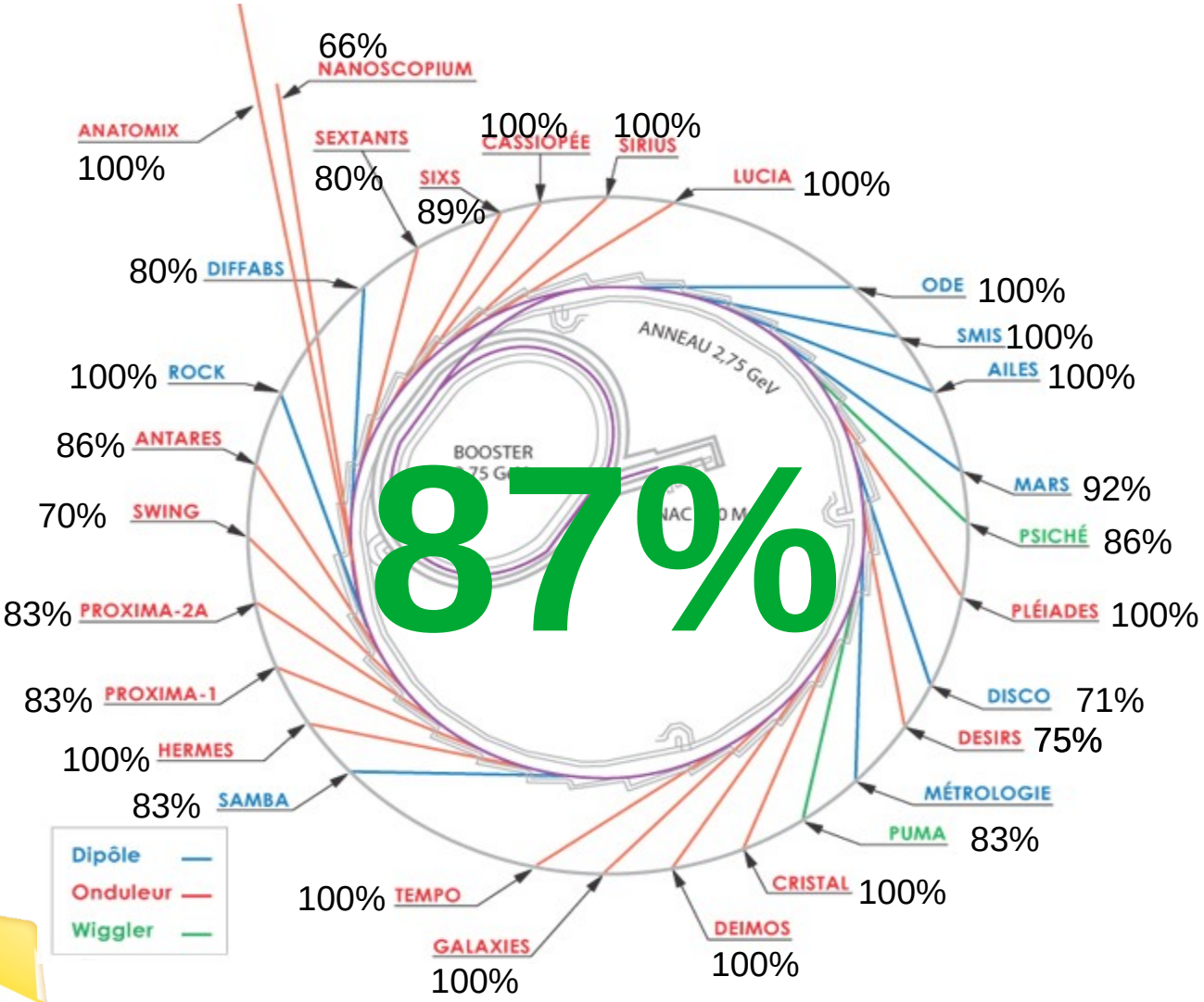
SOLEIL II
La science éclaire l'avenir
Science lights up the future

Data reduction and Analysis Group at SOLEIL

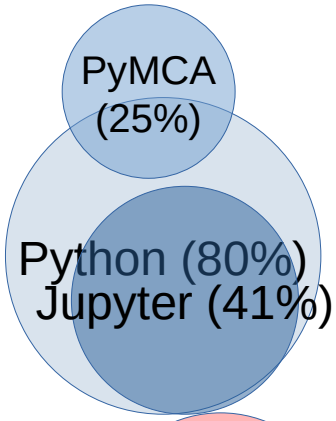
- **G**roupe de **R**éduction et d'**A**nalyse de **D**onnées **E**xpérimentales de **S**oleil (GRADES).
- Created in Feb 2020.
- Staff: 6 perm. (3 dev, 2 BL, 1 AI) + ~4 temp.
  - E. Farhi, F. Picca; O. Roudenko, S. Bac; A. Bellachehab, G. Cassam
- Topics: scientific software (diff, spectro/abs, simulation, ...), Debian/Ubuntu packaging, deployment, coding, documentation, support, … and now AI everywhere.
- Means: 3 central servers + ~20 BL servers.

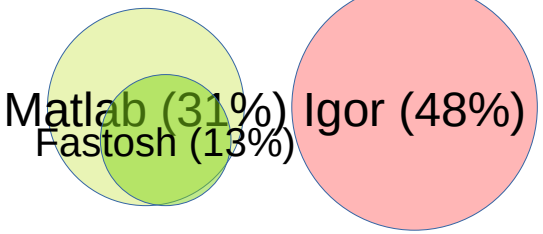
SOLEIL II
La science éclaire l'avenir
Science lights up the future

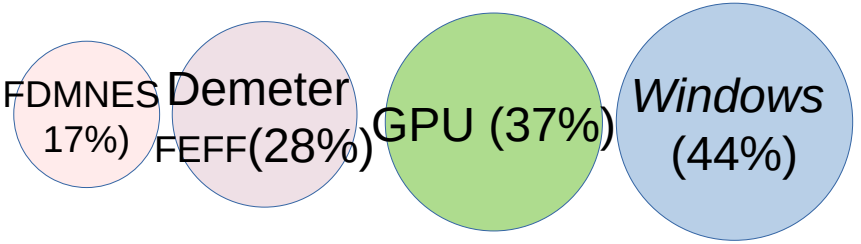
Beam-line requirement capture



29 BLs
~100 “crazy” scientists
160 requirements
144 supplied (87%)







Beam-line requirements: software

Our BLs require essentially :

- **Quasar/Opus PyHST2 UFO** ImageJ Fiji **Napari Nabu Astra-Toolbox** *matlab* **silx navarp arpys binoculars pynx** labplot **jupyterhub jupyterlab** codeblock vscode **pyfai pymca** maud **athena/artemis** correli-q4 cryspy/quanty fdmnes quantum-espresso/**sssp** abinit cp2k gromacs lammps fastosh **xraylarch** sasview **cctbx dioptas bornagain genx** anarod **xrayutilities** ase foxtrot atsas **bioxtas-raw** matlab-csaxs-ptycho
- *Igor RefFIT casaxps* (Windows)

We develop specific software:

- **binoculars/hkl**, **McXtrace**, **pyfai/pynx**, **qemu-web-desktop**, notebooks, BL data I/O, ...

Software handling

- We have chosen **Debian-class** systems to ensure coherence, stability and maintainability of most of our scientific software.
  - The Debian Science stack is extensive.
  - Brings better life-span than **conda** and **containers**.
  - We use ‘unstable’ packages on request.
- Two Debian contributors in the group + 40k€/y subcontracting.

  ~**400** packages pushed in Debian Science over 4 years.

  https://salsa.debian.org/pan-team/soleil-packaging-overview

F. Picca
E. Farhi
R. Mas
X. Guimard

Data reduction

Load raw data (handle formats).

Normalize/correct intensity:

$$\frac{Light - Dark - Offset}{Flat - Dark - Offset}$$

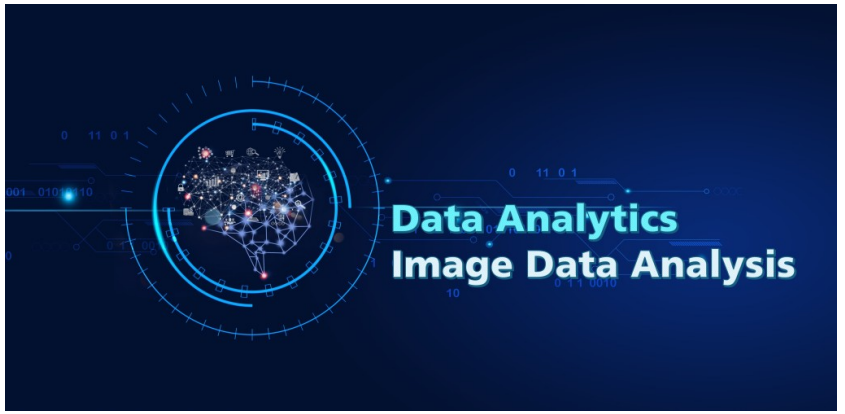
Align (image registration) and **Stack**.

Transform coordinates. Add metadata.

We mostly use Python libraries:

- `pyfai, pymca, silx, h5py, libxrl, xraydb, larch, pyhst/nabu, skimage, sklearn, orange3/quasar, pynx, …`

Data analysis



Load reduced data.

Make use of **metadata**.

Use **large data sets**, #CPU and #GPU.

Usually perform **statistics**.

Fit theory (extract physical parameters).

Estimate **errors**.

Write scientific **papers**.

- `Dioptas, Mantis-Xray, Larch, Arpys/Navarp, Orange3/Quasar, BioXtas-RAW, SASView, BornAgain, GenX, Napari, Fiji, …`

Data analysis services

We provide all inclusive  Data Reduction, Analysis and Visualization services:

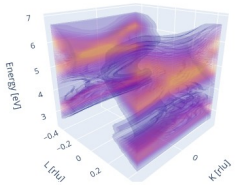
Thousands of software and libraries.

DARTS (easy install/deploy).

https://data-analysis.synchrotron-soleil.fr/desktop

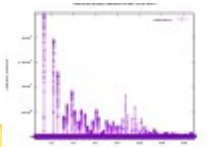
JupyterHub (shared storage with *DARTS*).

https://data-analysis.synchrotron-soleil.fr/notebook/



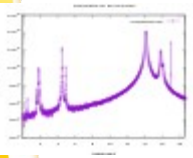
Material modelling (easy DFT+e^-/*ph* dispersions)

https://data-analysis.synchrotron-soleil.fr/computing/material_modeling.html



Diffraction + Fluorescence estimates

McXtrace

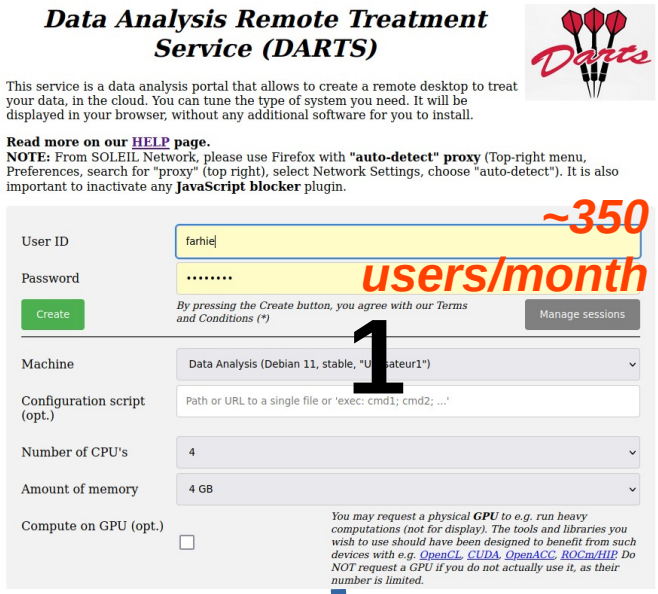


https://data-analysis.synchrotron-soleil.fr/computing/fluo_diffraction.html

SOLEIL II
La science éclaire l'avenir
Science lights up the future

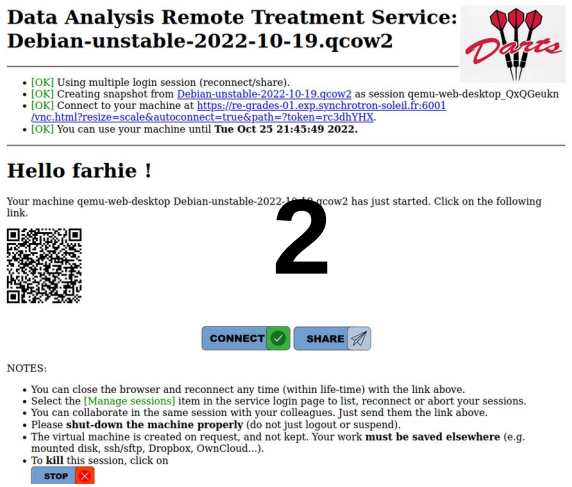
Data analysis: DARTS

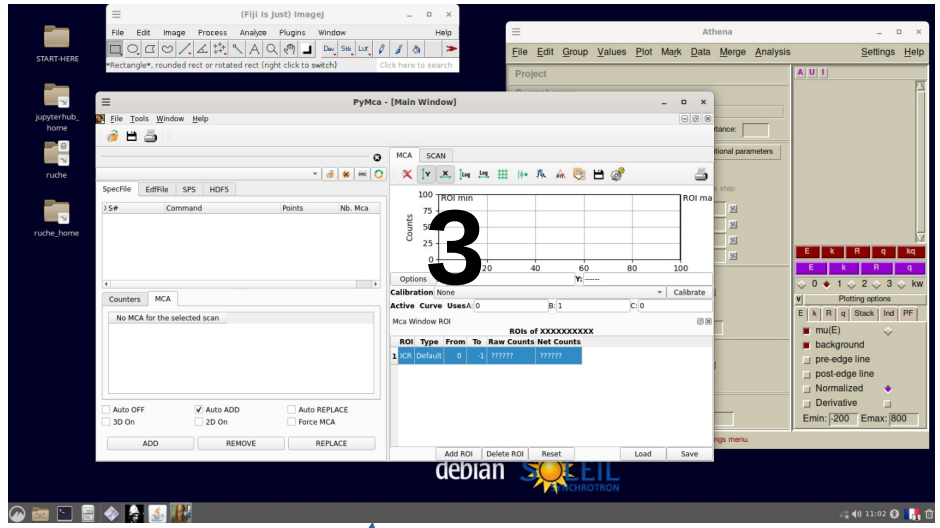
Service: https://data-analysis.synchrotron-soleil.fr/
Code: https://gitlab.com/soleil-data-treatment/soleil-software-projects/qemu-web-desktop



Currently Includes:

- CP2K, GROMACS, LAMMPS, QE/SSSP, ABINIT, ASE, Jupyter NB, FEFF/FDMNES, ORCA, FPLO, PyTorch, Sk-Learn/image...





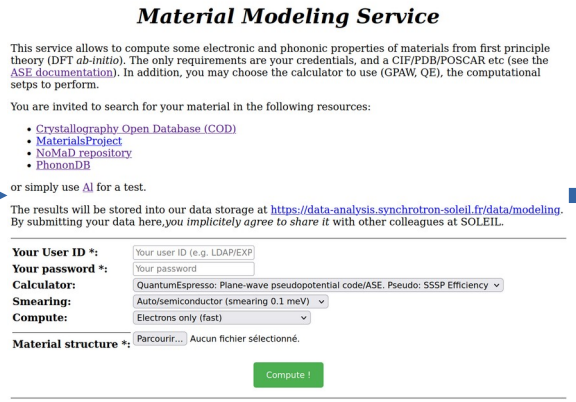
DARTS is simple and efficient.
Fully configurable.
Low Tech / Maintenance.
Deployment in **minutes**.
Scalable.

QR code.
Access to **GPUs**.
Open-source/Debian.
100+ of sci software in VM.
Access to a persistent area.
Access to the Ruche (exp data).
Will evolve into the **VISA** service.

Material modelling (easy DFT+e^-/*ph* disp.)

- **Service:** https://data-analysis.synchrotron-soleil.fr/computing/material_modeling.html
- **Code:** https://gitlab.com/soleil-data-treatment/soleil-software-projects/material-modelling-service

Input
CIF



Material Modeling Service

This service allows to compute some electronic and phononic properties of materials from first principle theory (DFT *ab-initio*). The only requirements are your credentials, and a CIF/PDB/POSCAR etc (see the ASE documentation). In addition, you may choose the calculator to use (GPAW, QE), the computational setps to perform.

You are invited to search for your material in the following resources:

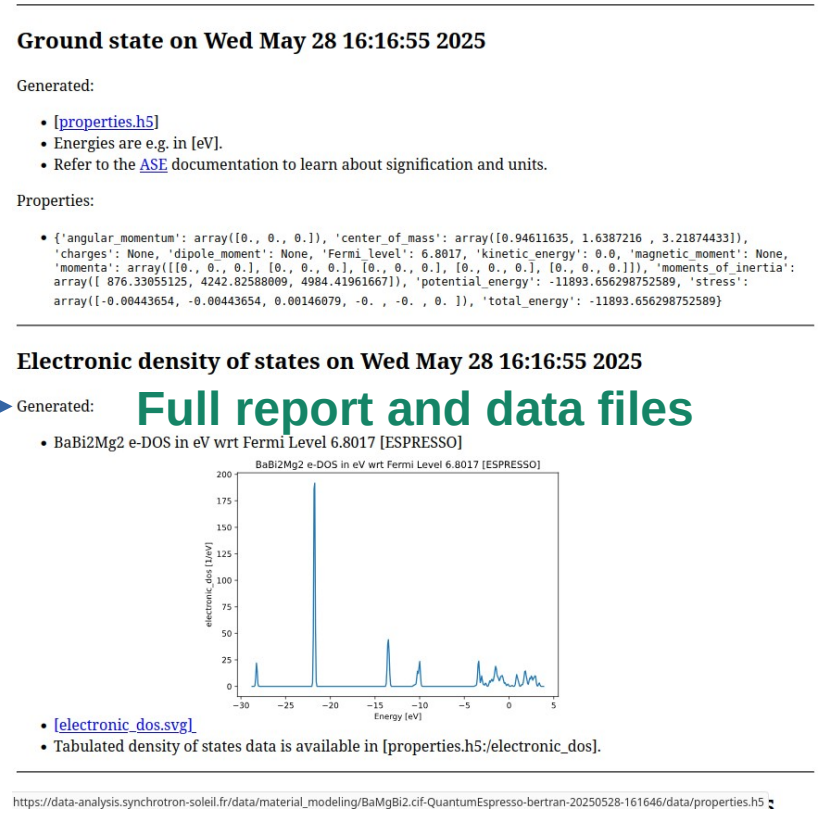
- Crystallography Open Database (COD)
- MaterialsProject
- NoMaD repository
- PhononDB

or simply use AI for a test.

The results will be stored into our data storage at https://data-analysis.synchrotron-soleil.fr/data/modeling. By submitting your data here,*you implicitely agree to share it* with other colleagues at SOLEIL.

Your User ID *: Your user ID (e.g. LDAP/EXP)
Your password *: Your password
Calculator: QuantumEspresso: Plane-wave pseudopotential code/ASE. Pseudo: SSSP Efficiency
Smearing: Auto/semiconductor (smearing 0.1 meV)
Compute: Electrons only (fast)
Material structure *: Parcourir... Aucun fichier sélectionné.

Compute !



Ground state on Wed May 28 16:16:55 2025

Generated:

- [properties.h5]
- Energies are e.g. in [eV].
- Refer to the ASE documentation to learn about signification and units.

Properties:

- {'angular_momentum': array([0., 0., 0.]), 'center_of_mass': array([0.94611635, 1.6387216 , 3.21874433]), 'charges': None, 'dipole_moment': None, 'Fermi_level': 6.8017, 'kinetic_energy': 0.0, 'magnetic_moment': None, 'momenta': array([[0., 0., 0.], [0., 0., 0.], [0., 0., 0.], [0., 0., 0.], [0., 0., 0.]]), 'moments_of_inertia': array([876.33055125, 4242.82588009, 4984.41961667]), 'potential_energy': -11893.656298752589, 'stress': array([-0.00443654, -0.00443654, 0.00146079, -0. , -0. , 0.]), 'total_energy': -11893.656298752589}

Electronic density of states on Wed May 28 16:16:55 2025

Generated:

Full report and data files

- BaBi2Mg2 e-DOS in eV wrt Fermi Level 6.8017 [ESPRESSO]

- [electronic_dos.svg]
- Tabulated density of states data is available in [properties.h5:/electronic_dos].

https://data-analysis.synchrotron-soleil.fr/data/material_modeling/BaMgBi2.cif-QuantumEspresso-bertran-20250528-161646/data/properties.h5

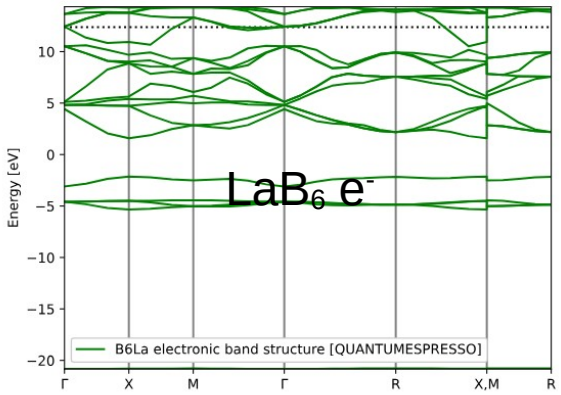
SOLEIL II
La science éclaire l'avenir
Science lights up the future

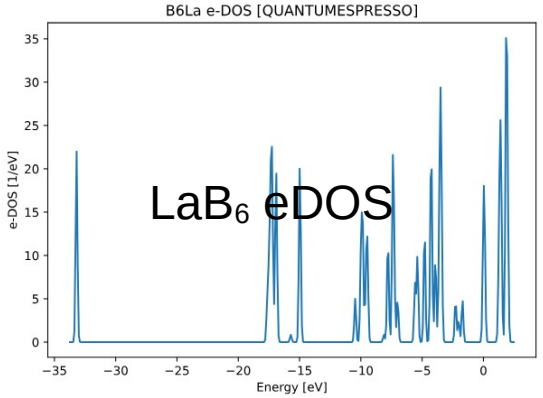
Material modelling (easy DFT+e^-/*ph* disp.)

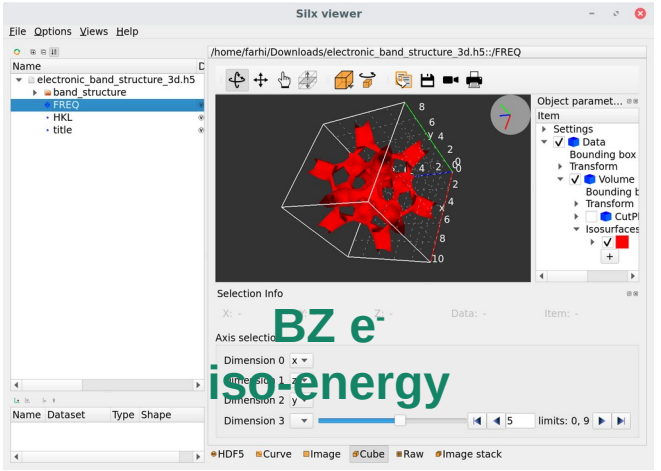
- We compute the electronic levels in the BZ using e.g. *QE/SSSP* via *ASE*.
- Configuration is automatic for fast estimates (based on *Materials Project* settings):

Calc	Conv. dE	Displ. δx	E_{cut}	Kpts	supercell
QE/SSSP	1e-5 eV	0.01 Angs	200*N_{at} eV	Nkpts/ $\sqrt[3]{(N_{at}*supercell)}$	$\sqrt[6]{(Kgrid^3/N_{at})}$

- From this we get electronic dispersions:







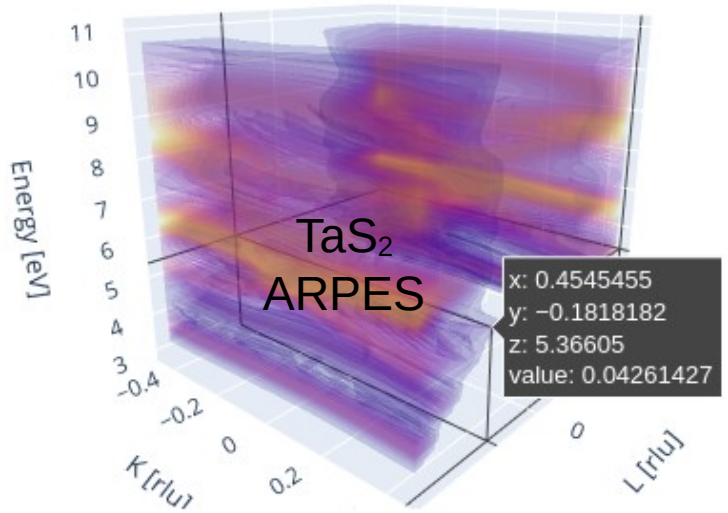
Electronic states → ARPES intensity

- From the Fermi golden rule via the electro-magnetic field ***A*** and light momentum ***p***

$$\Gamma_{i\to f} \propto |<f|A.p|i>|^2 \delta(E_f - E - h\nu)$$

- Field ***A*** boils down to photon polarization, and ***p*** to the in-plane momentum (surface).
- We compute an estimate for the intensity on a flat sample surface

$$I(q,E) \propto \sum_{bands} |\vec{\epsilon}.\vec{q}_{||}|^2 \frac{1}{1+e^{(E-E_F)/kT)}} \frac{1}{1+(E-E_f)^2/\Gamma^2}$$



Small displacements → vibrational states

- From the Fermi golden rule via the interatomic potential *V* and particle density ρ.

$$w_{\vec{k}_i \to \vec{k}_f, \lambda_i \to \lambda_f} = \frac{2\pi}{\hbar} \left| \langle \lambda, \vec{k}_f | V | \lambda_i, \vec{k}_i \rangle \right|^2 \rho(E_f),$$

Use *PhonoPy*

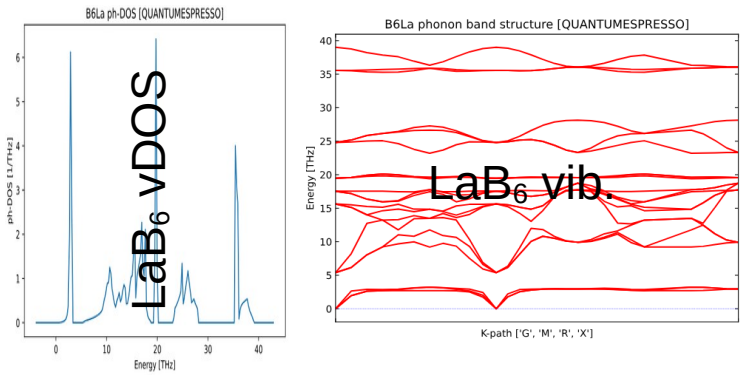
- We compute the form factors (harmonic approx – small displacements → forces and potential *V*)

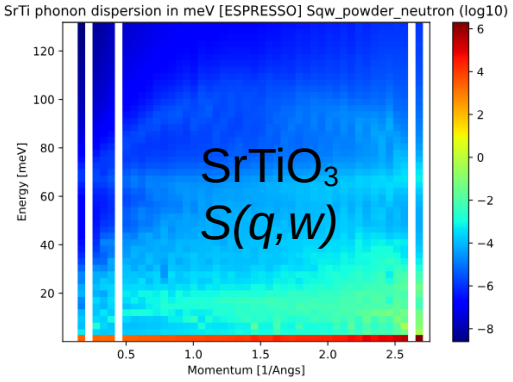
$$F_j(Q) = \sum_a \frac{b_a}{\sqrt{m_a}} e^{-W_j(Q)} \vec{Q}.\vec{e}_j e^{-i\vec{Q}.\vec{r}} \qquad W_j(Q) = \frac{|\vec{Q}.\vec{e}_j|^2}{2\omega_j(Q) m_a} (2n(\omega_j(Q) + 1) \qquad n(\omega) = 1/(e^{\hbar\omega/k_B T} - 1)$$

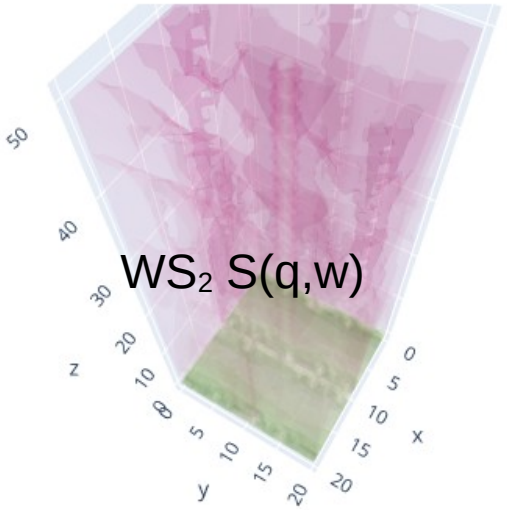
b = scatt. length [m] = r_e *f* for photons

- And then the X-ray and neutron vibrational dynamic structure factors.

$$S(\vec{Q}, \omega) = \sum_j (n(\omega) + 1) \frac{|F_j(Q)|^2}{\omega_j} DHO(\omega, \omega_j, \Gamma)$$







Material fluorescence modelling service

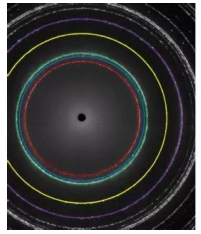
Diffraction + Fluorescence estimates

- https://data-analysis.synchrotron-soleil.fr/computing/fluo_diffraction.html
- https://gitlab.com/soleil-data-treatment/soleil-software-projects/fluo-diffraction-service

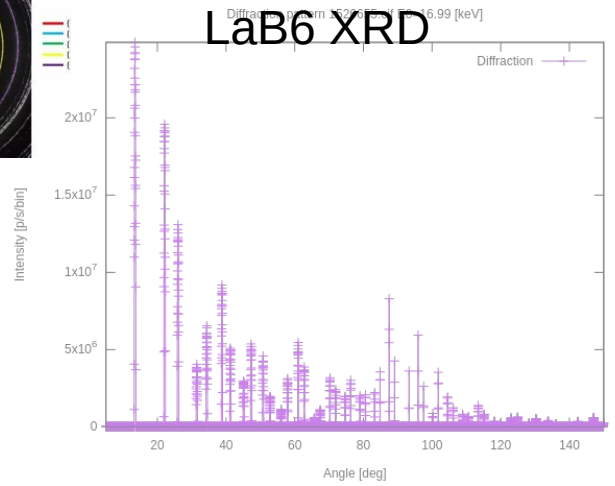
We rely on XrayLib (librxl) and McXtrace to compute the powder diffractogram and the fluorescence spectra. Also compute the experimental “background”. Produce ***realistic*** data sets.

$$\sigma_{\mathrm{cone,Q}} = j_Q f \exp(-2W)\frac{V}{V_0^2}\frac{\lambda^3}{4\sin\theta}|F(Q)|^2$$

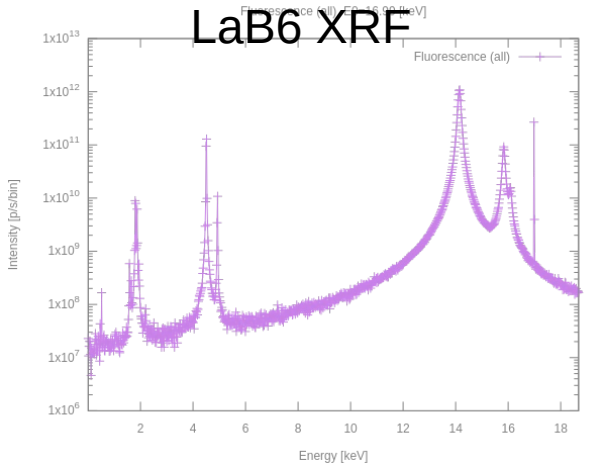
$$I_{\mathrm{el}} = I_{\mathrm{o}}\frac{\Omega}{4\pi}\frac{\sigma_{\mathrm{XRF}}(E_0)}{\sigma_{\mathrm{TOT}}(E_0)}\frac{1}{\sin\psi_1}\frac{C_{\mathrm{el}}\,\alpha_{\mathrm{el}}(E_0)}{\frac{\alpha(E_0)}{\sin\psi_1}+\frac{\alpha(E_1)}{\sin\psi_2}}$$



LaB6 XRD



LaB6 XRF



McXtrace

https://mcxtrace.org/

E. Knudsen *et al,* JAC **46**(3):679-696, 2013
A. Brunetti *et al.*, Spectrochimica Acta Part **B 59** (2004) 1725–1731
T. Schoonjans *et al.,* Spectrochimica Acta Part **B 66** (2011) 776–784

Material Databases

Before starting any new study from scratch, you should consider to inquire for existing structures and vibrational information from **databases** at:

- https://alexandria.icams.rub.de/
- https://optimadeclient.materialscloud.io/ (catalog of catalogs)
- https://www.materialscloud.org/home (QE team)
- https://nomad-lab.eu/nomad-lab/ (simulated structures)
- https://next-gen.materialsproject.org/ (simulated structures)
- https://www.crystallography.net/cod/ (sim and exp)
- https://github.com/atztogo/phonondb (phonopy db)

A minute of silence

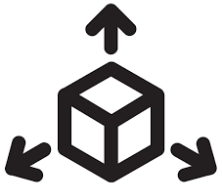
The advent of A.I.
The end of equations and theory.

The “A.I.” altar

- A.I. algorithms mostly rely on “**training**”, and store vast amounts of data.
- The data must be properly “**annotated**”.
- The A.I. connects/correlates “inputs” with “outputs”. It “mimics” the training data: a **parrot**.

- The logical path that produces the A.I. answer (aka *inference*) is not always explicit.

  It behaves as an oracle (*no critics here*): a **black box**.

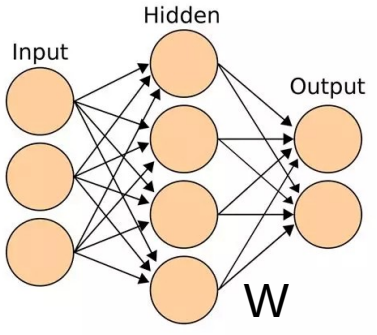


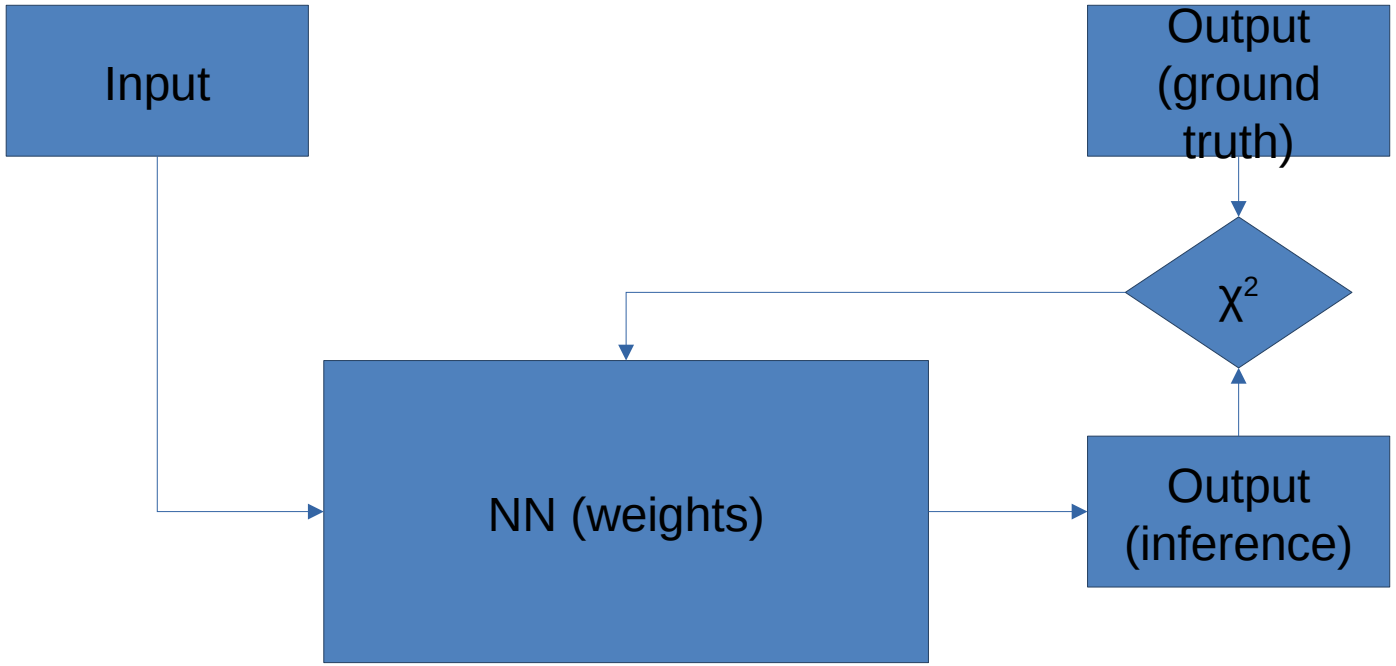
- The A.I. is **restricted** to the training data range.
- The **“accuracy” is hard to estimate**, so that A.I. does NOT follow the science path.

Conventional neural networks

We have an input, an output, and some internal ‘weights’. We optimize the NN internal weights (the ‘memory’) so that its output converges to the expected one.

This is a **pure optimization algorithm**.
The large number of parameters allows to approximate any function.





Looks like an Ising interconnected spin-glass

Transformers / Self-Attention

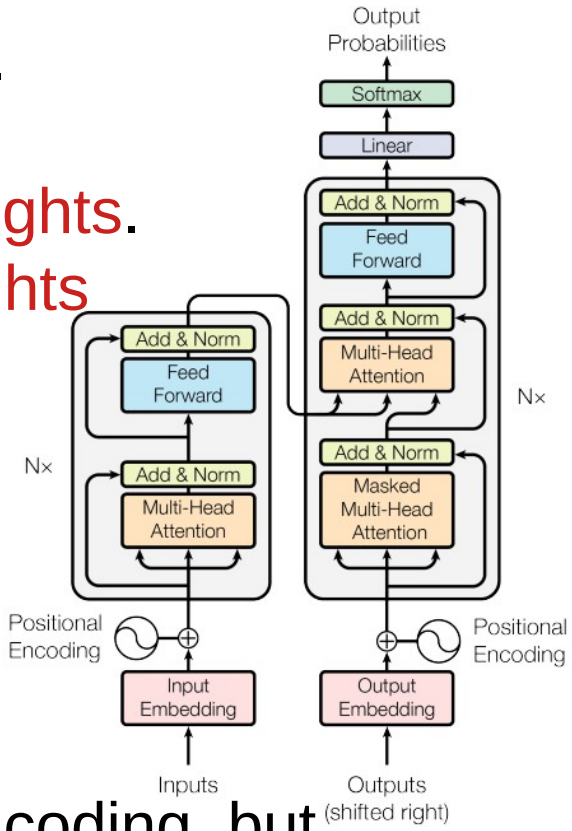
Transformers rely on pure **linear algebra**.

We first 'embed' the input as *tokens* ***x*** (via quantization and PCA).
The output of the Transformer has the same size as the input ***x***.
The knowledge is stored into some Queries, Keys and Values weights.
Q.K.V vectors are determined from embeddings and internal weights

$$QKV = x.W_{QKV}.$$

The weights are set during training.
The output is a dot product:

$$y = softmax(Q.K^T/\sqrt{d}).V$$ [softmax = logistic = Boltzmann]

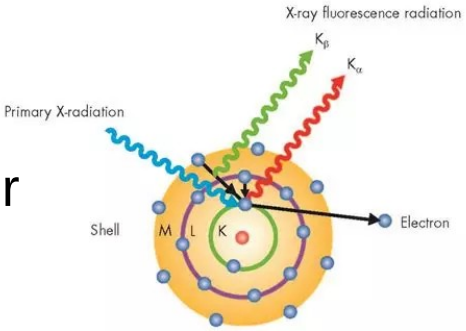


Transformers can NOT be exact as they rely on reduced latent encoding, but they are (nearly) deterministic.

See: https://xmarva.github.io/blog/2025/building-a-transformer/

Automatic data reduction. Where ?

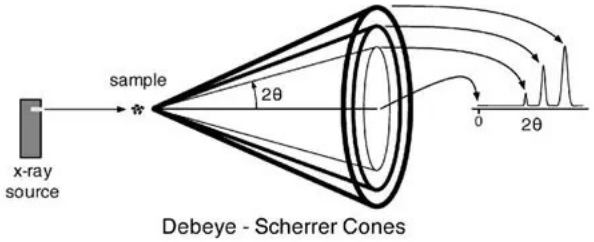
All beam-lines are equipped with **fluorescence** detectors



- Photon ejects an electron in an absorption process
- The hole is filled by an electron that “falls down” while emitting a photon at lower energy
- Generates a 1D energy spectra (e.g. 1-40 keV)

Many beam-lines are equipped with **diffraction** 2D detectors

- The crystalline periodic structure diffracts with Bragg law (coherent elastic scattering)
- Creates a 2D set of “rings” or “spots” *vs* angle



Debeye - Scherrer Cones

Other possible techniques

- **absorption edges**: specific to atoms oxidation and coordination
- **small angle scattering**: specific to “large” structures (micelles, colloids, polymers, proteins, …)
- Multipurpose **chat**

Is A.I. suitable for science ?

- A.I. can be used for **automatic interpretation/reduction**.
- **We have started:**
  - ➔Language models → Help scientists
  - ➔Fluorescence/XRF → Elemental analysis
  - ➔Diffractogram/XRD → Space-group
  - ➔SAXS → Large-structure model (SASView)
  - ➔XAS → Oxidation/coordination

We aim to provide cloud services for such X-ray applications.

AI@SOLEIL language models

- **Warning**: keep property of your data

  avoid cloud services

- **Warning**: how do you estimate if answer is valid ?

  trust your own expertise – just as you would do on Internet

- **What for**: help at 3pm, during acquisition or when preparing proposals, broad tech advice.

- **We develop local LLMs**

  Mistral/DeepSeek in production as a local service : we use **llama.cpp** as inference engine and web UI

  Runs on two servers, with ‘thought’ process.

  Mistral AI connected to our **Confluence** knowledge base (100 % *original work*)

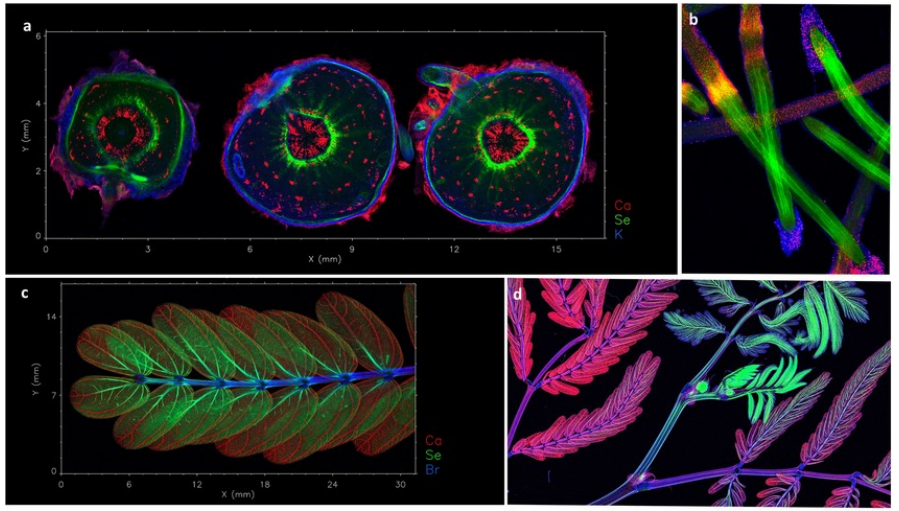
  Context tuning for a RAG service (upload doc and/or search local doc)

Started end 2023
J. Hassani
A. Bellachehab

AI@SOLEIL fluorescence

Expected service

- **IN**: raw fluorescence data (binned on 0-45 keV range)
- **OUT**: atoms and stoichiometry (Z=5-90). Elemental analysis
- (opt) Estimate for incident energy
- (opt) Estimate for detector artifacts



Usable as script, and web API

- Fast estimates for any 1D, 2D, 3D data with fluorescence spectra
- Fast guess for e.g. *pymca* and *xray-larch*.
- Results stored as HDF5/NeXus metadata
- Should be ready end of 2025

Started March 2025
F. La Porta
S. Bac
A. Bellachehab

AI@SOLEIL fluorescence

- We have started to train an A.I. to analyze spectroscopy data.

⚠There are no massive annotated data sets.

Don’t panic, this is usual with A.I.

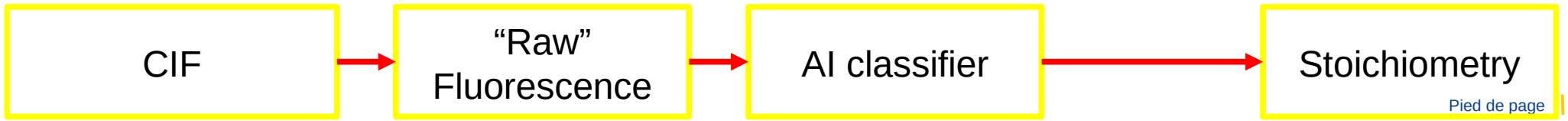
- We use 600k compounds and associate fluorescence spectra.

  Assemble a **McXtrace** model of a fluorescence X-ray imaging beam-line.

  Contains photon source, optics, sample (*Fluorescence*) and detectors.

    Include multiple-scattering, Compton, Rayleigh, fluorescence cascade, absorption secondary extinction, sample geometry, detector artifacts, Monte-Carlo noise as in experiments, ...

    Relies on **xraylib** (ESRF): Boron to Thorium (Z=5-90)



Beam-line modelling

McXtrace is a general Monte Carlo ray-tracing software for simulation X-ray beamlines and experiments.

https://mcxtrace.org/

Forked from **McStas** for neutron scattering

www.mcstas.org

Provides

209 components (sources, optics, samples, monitors, etc)
98 examples

The strength and specificity of McXtrace and McStas is the ability to model sample scattering within instruments/beam-lines.

Project funded by EU since 1996. Mostly C-based, with python interfaces.

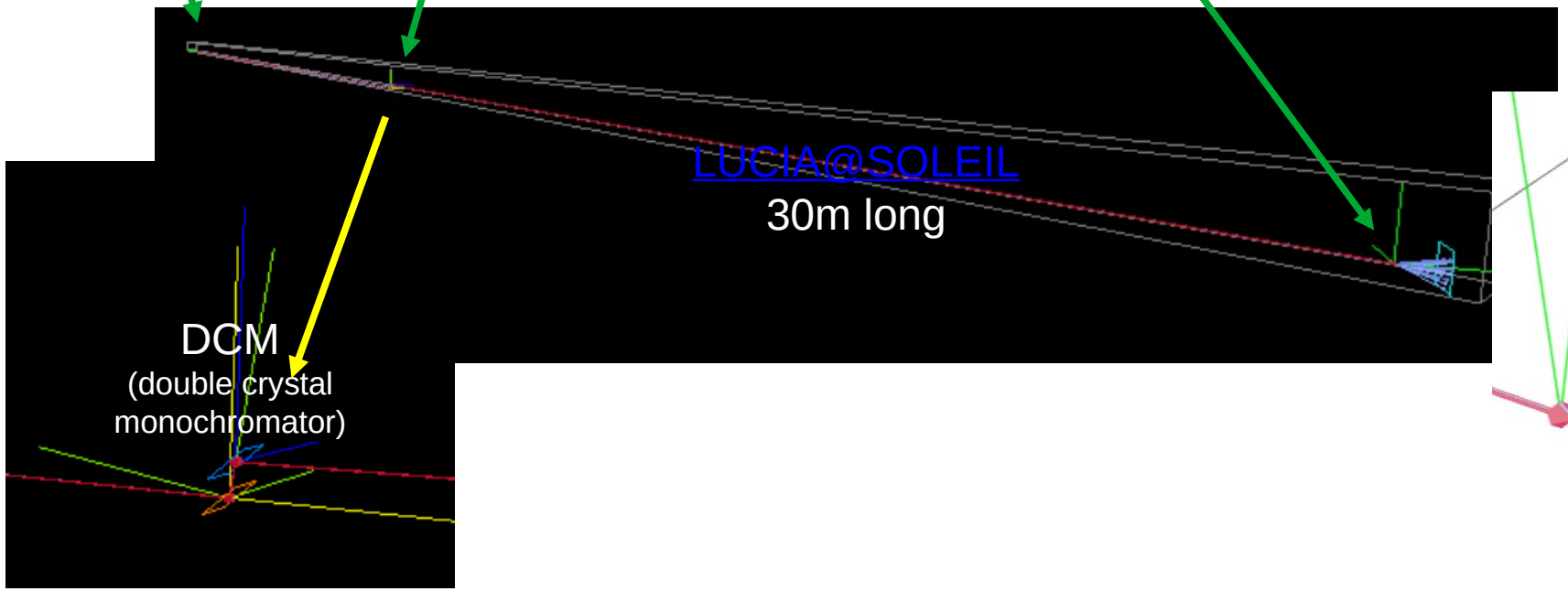
https://github.com/mccode-dev/McCode

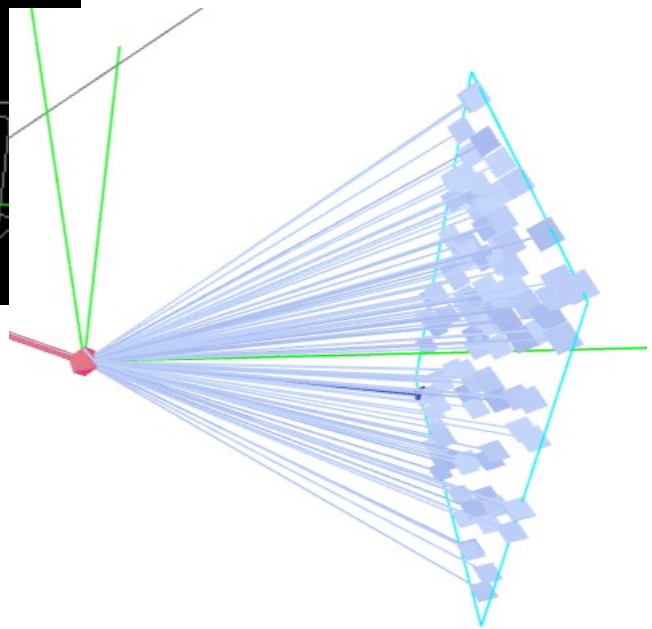
Runs on all architectures (conda, debian-class, …). Massively parallel: MPI, GPU

Generating realistic annotated training set

McXtrace

```
COMPONENT source = Undulator(E0 = E0, dE = dE, …)
        COMPONENT dcm_xtal0 = Bragg_crystal(length=0.04, width=0.04, h=1, k=1, l=1, material="Si.txt", V=160.1826)
        COMPONENT dcm_xtal1 = COPY(dcm_xtal0)
                COMPONENT sample = Fluorescence(material=sample_material…)
                COMPONENT mon_spl_fluo = Monitor_nD(options="energy", ...)
```





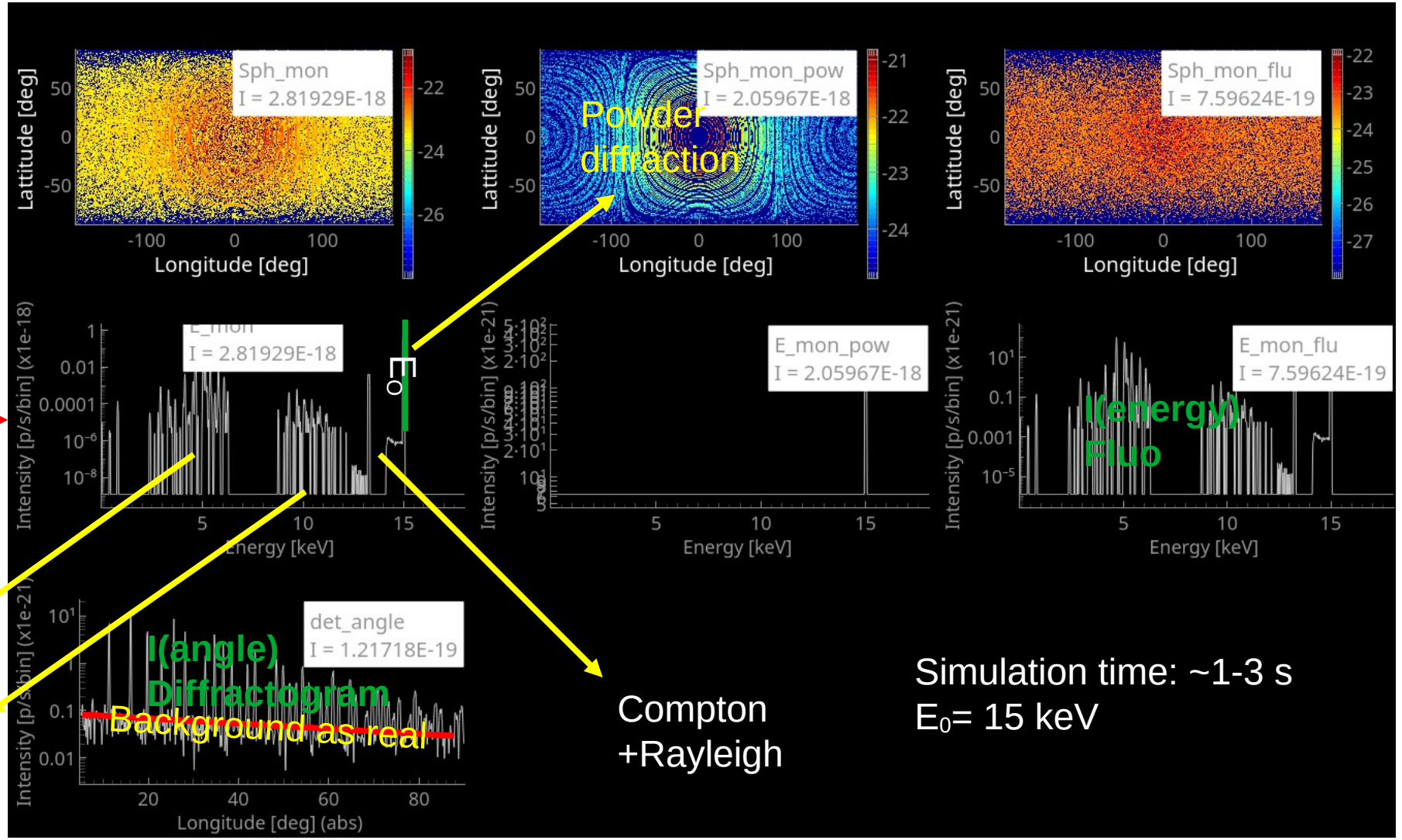
SOLEIL II
La science éclaire l'avenir
Science lights up the future

Generating realistic annotated training set

Using FluoPowder
McXtrace sample

Fluorescence is not really isotropic – appears as a background in angle

CIF (structure)



Fluorescence

SDD pile-up
SDD escape

Compton
+Rayleigh

Simulation time: ~1-3 s
E_0= 15 keV

AI@SOLEIL fluorescence: using Materials Project

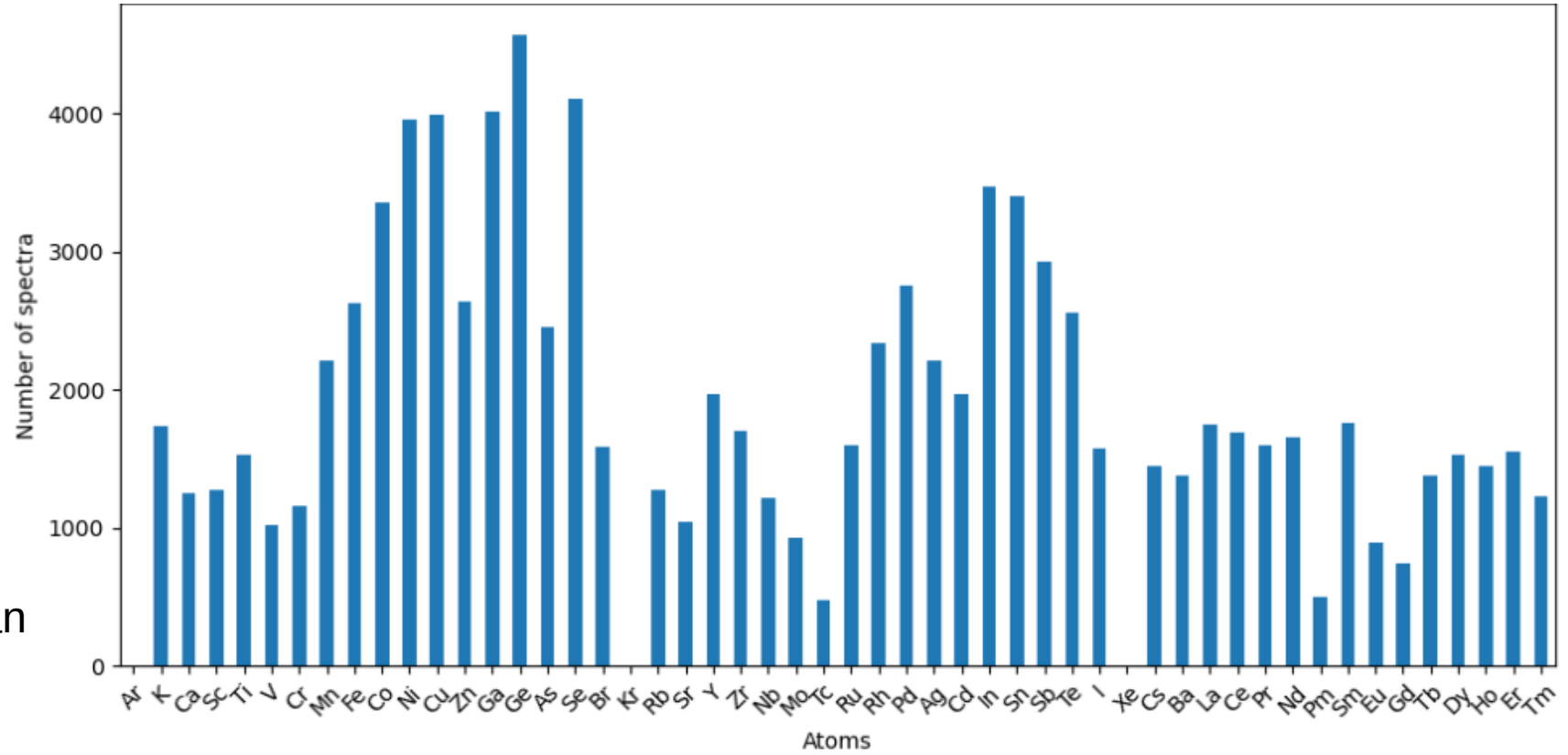
Spectra simulation

Simulation model

- Gaussian x-ray sources
- Sample

Simulation

E_0 = 60 keV
Range 0-66 keV
n=10^5 rays
Spectra with 2000 points
Fluorescence line set as Gaussian



Formula

- Real formula containing at least 1 element with Kα between 3keV < E_{K_α} < 60keV (52 elements)
- Formula from MP, formula containing only that 52 elements
- 34136 spectra

AI@SOLEIL fluorescence: energy encoding

Positional encoding for LLM

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right); PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

i: dimension index
d: embedding size

Custom made positional encoding for spectroscopy

$$PE(E, 2i) = \sin\left(\frac{E}{\tau^{\frac{2i}{d}}}\right); PE(E, 2i+1) = \cos\left(\frac{E}{\tau^{\frac{2i}{d}}}\right)$$

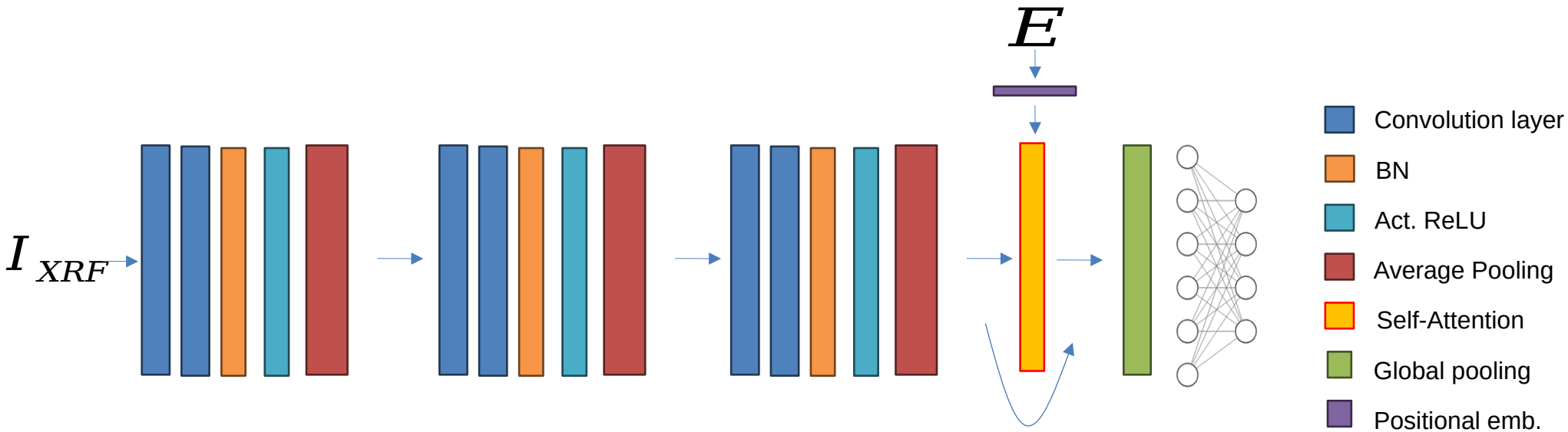
To get $\tau \rightarrow$ PE must deliver unique values until 66 keV : $\sin(E_{max}/\tau) = \sin(2\pi) \rightarrow \tau = 66/2\pi \sim 10$

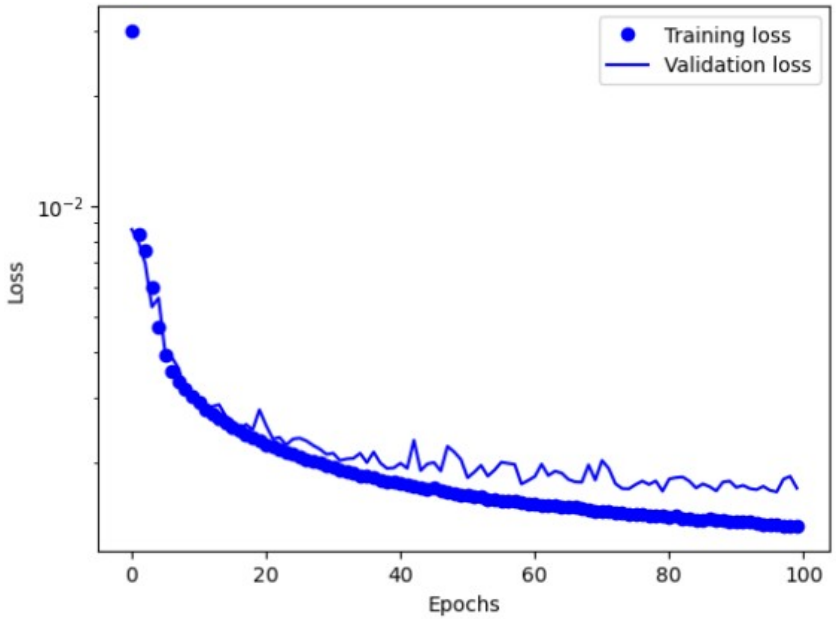
$$PE(E, 2i) = \sin\left(\frac{E}{10^{\frac{2i}{d}}}\right); PE(E, 2i+1) = \cos\left(\frac{E}{10^{\frac{2i}{d}}}\right)$$

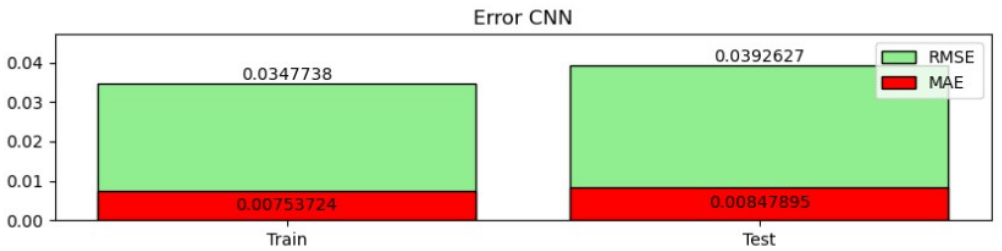
The positional encoder is summed to the embedding (feature maps of the conv)

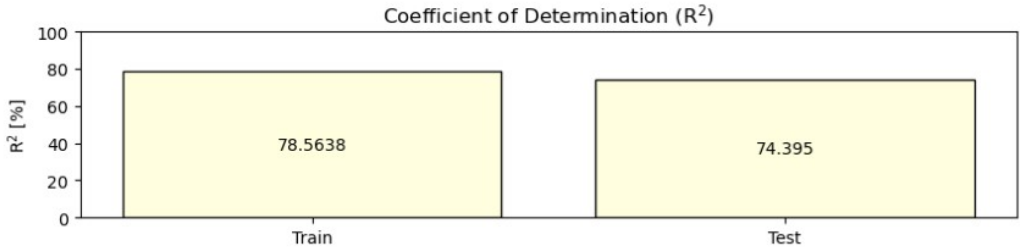
AI@SOLEIL fluorescence: model topology

- **Convolution layer 1**
  - Kernel size = 4
  - Filters: 16
  - Batch norm (BN)
  - Pool 2
- **Convolution layer 2**
  - Kernel size = 4
  - Filters: 32
  - Batch norm
  - Pool 4
- **Convolution layer 3**
  - Kernel size = 4
  - Filters: 34
  - Batch norm
  - Pool 4
- **Self Attention**
  - Attention head = 10
  - Dimension keys = 64
  - Skip layer
- **2x Fully Conn. layer**
  - 128 neurons
  - Activ. f. = relu
  - Dropout 10%
- **Ouput layer**:
  - sigmoid activation

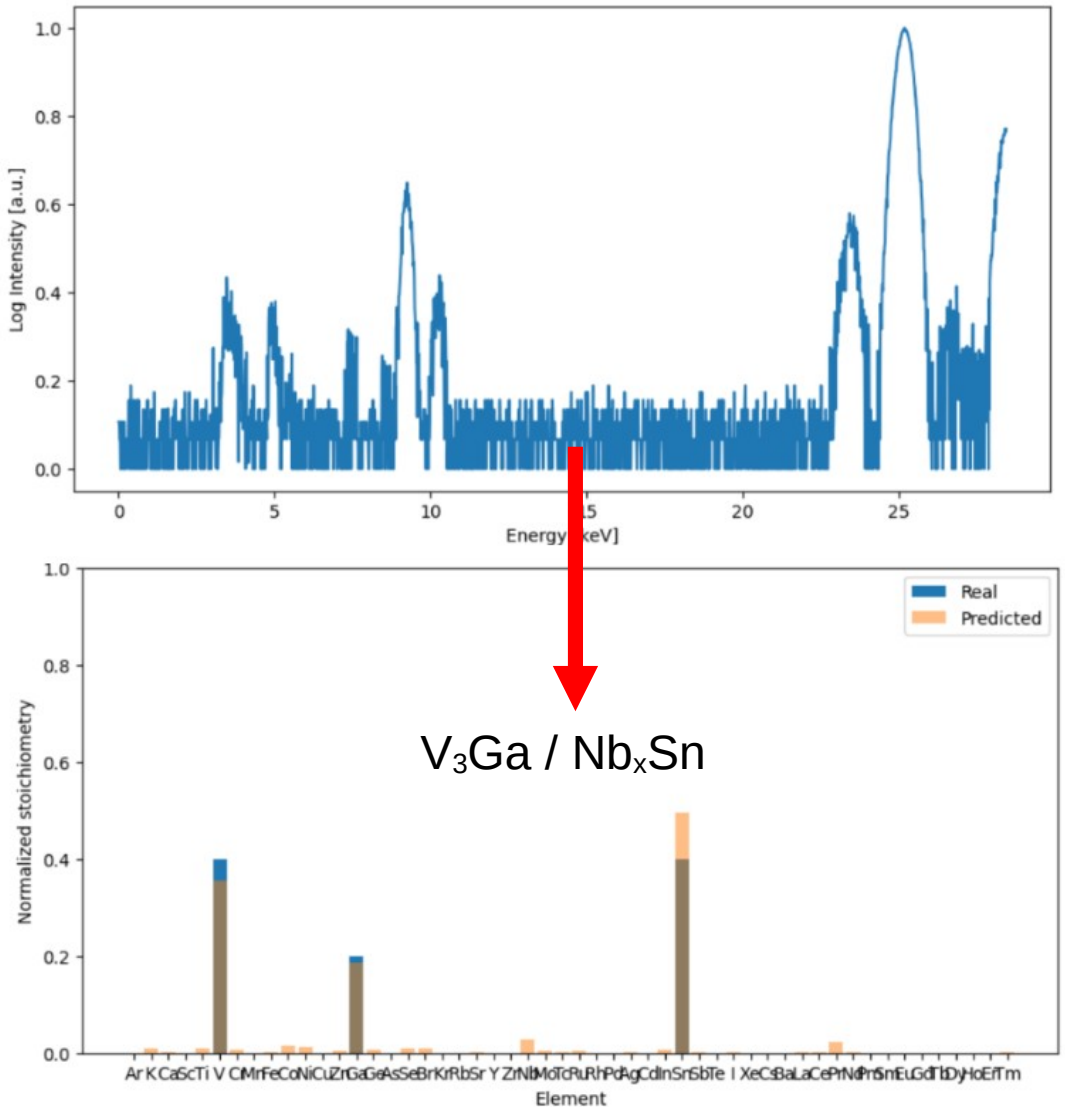






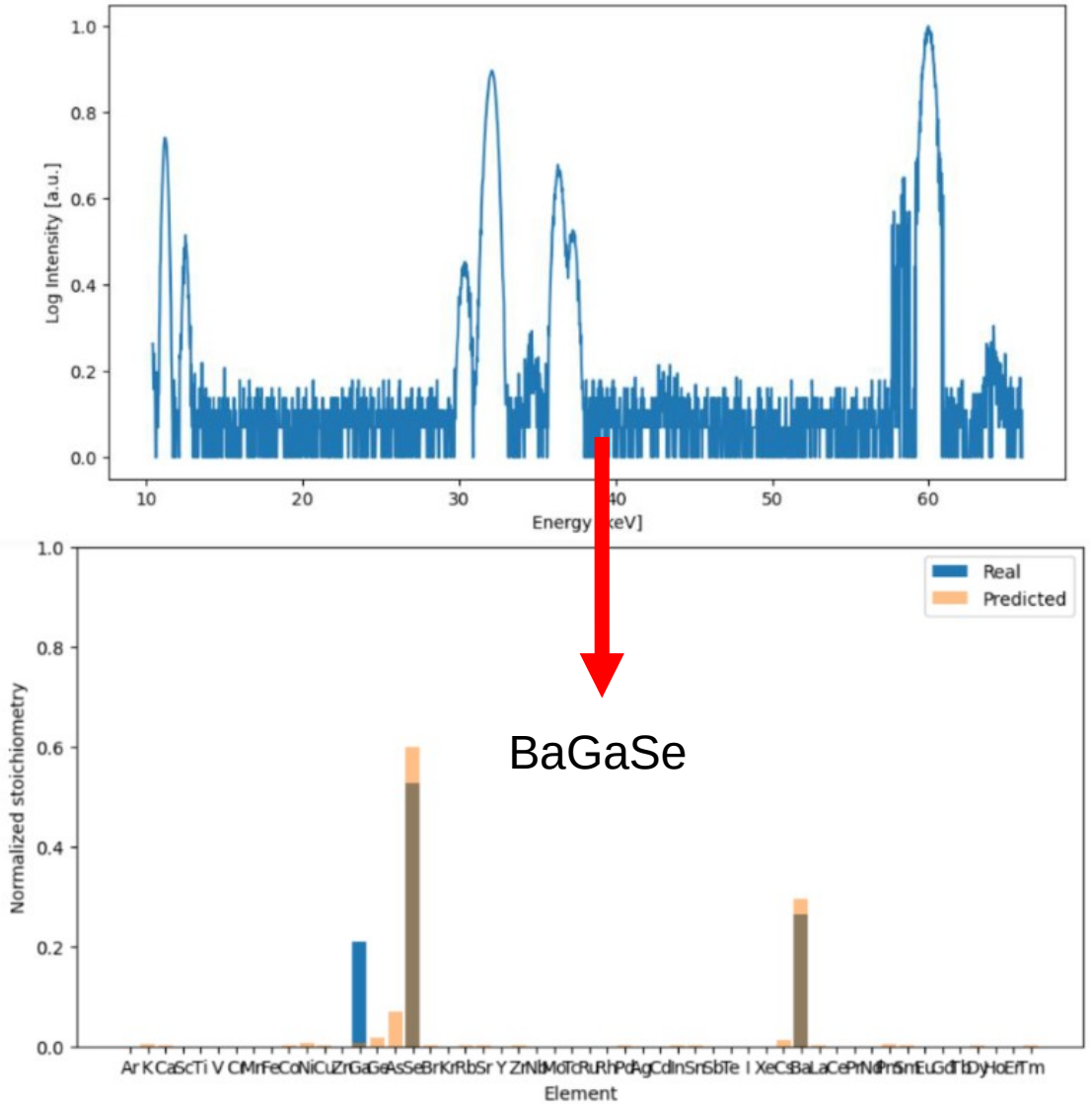


AI@SOLEIL fluorescence: results



V_3Ga / Nb_xSn

https://next-gen.materialsproject.org/materials/mp-22568?formula=V3Ga



BaGaSe

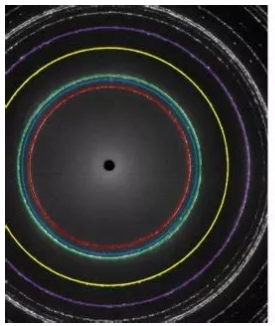
https://next-gen.materialsproject.org/materials/mp-1191072?formula=BaGa4Se7

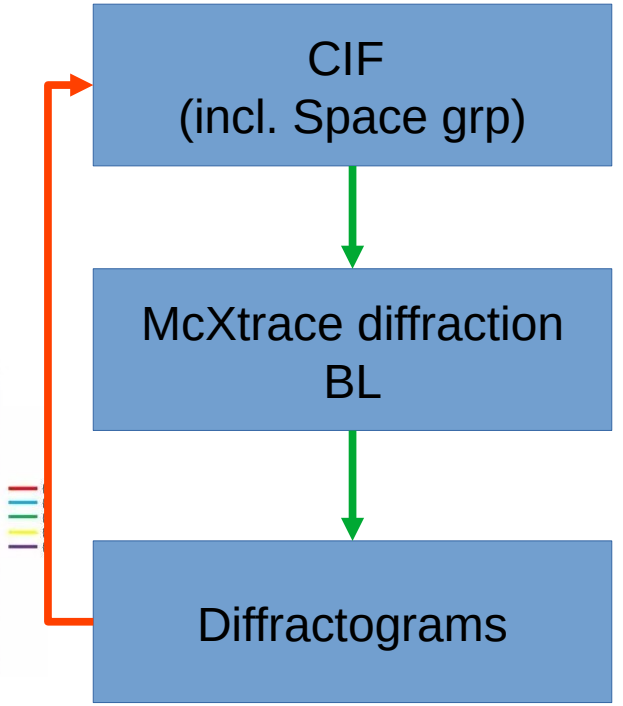
Next step: diffraction

- **IN**: raw diffractogram (binned on 0-180° range), mosaicity, fluo bkg.
- **OUT**: Space-group (1-230) and Bravais class

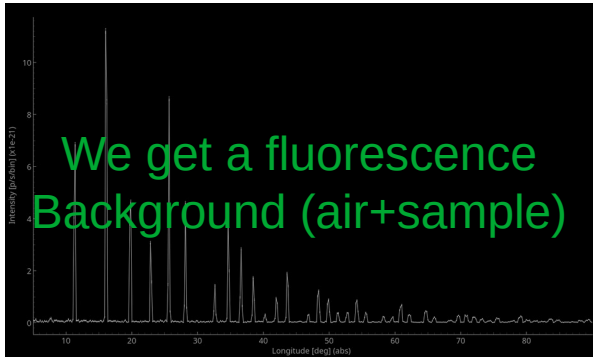
Usable as script, and web API

- Fast estimates for any 1D, 2D, 3D data with diffractograms
- Fast guess for e.g. FullProf, MAUD, TOPAZ, …
- Results stored as HDF5/NeXus metadata
- Probably also usable for neutrons
- Should be ready next year (but we already have a prototype)



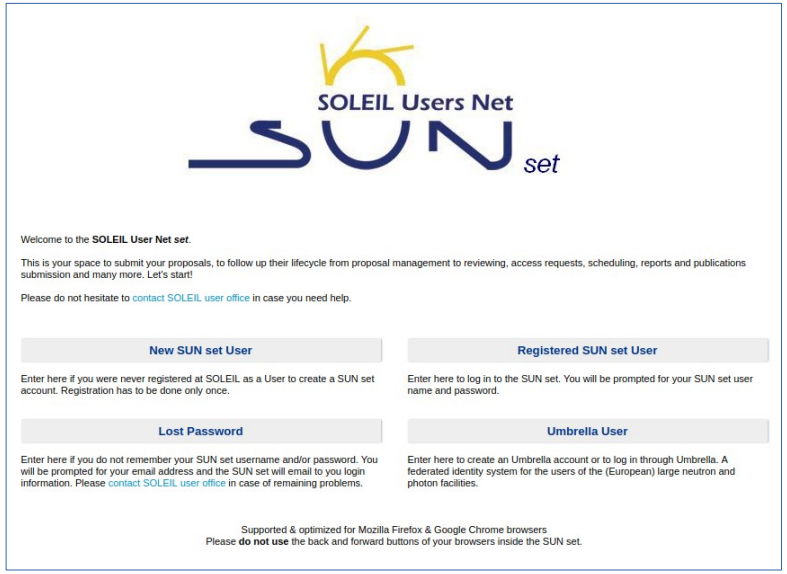


Prototype
A. Bellachehab



How to write a proposal for beam-time at SOLEIL

1) Specify the **sample** you wish to study.
2) Specify the type of **measurement** you aim for:
   - Diffraction is first required. If known, indicate structure in proposal.
   - ARPES and RIXS are well suited for solid-state physics.
3) Identify relevant target **beam-line**, and subcommittee panel:
   - Structure and elemental analysis : DIFFABS, CRISTAL, ...
   - Dynamics : CASSIOPEE, ANTARES, GALAXIES, …
4) Contact the BL **scientists** or <useroffice@synchrotron-soleil.fr>.
5) Write the scientific case, estimate measurement time (incl. Calibration and reference samples).
6) Justify the BL request (« *SOLEIL is the best ; that BL is really suited ; study never done before...* »)
7) Refine text in iterations with the BL scientist, well before the dead-line.
8) **Submit** at <https://sun.synchrotron-soleil.fr/sunset/bridge/sunset/>



Emmanuel FARHI <emmanuel.farhi@synchrotron-soleil.fr>