D5.4 – Report on lab-scale and large-scale facilities automated and standardized data analysis

VERSION

VERSION	DATE
1.0	15/11/2023

PROJECT INFORMATION

GRANT AGREEMENT NUMBER	957189
PROJECT FULL TITLE	Battery Interface Genome - Materials Acceleration Platform
PROJECT ACRONYM	BIG-MAP
START DATE OF THE PROJECT	1/9-2020
DURATION	3 years
CALL IDENTIFIER	H2020-LC-BAT-2020-3
PROJECT WEBSITE	big-map.eu

DELIVERABLE INFORMATION

WP NO.	5
WP LEADER	CEA, Sandrine Lyonnard
CONTRIBUTING PARTNERS	ILL, CEA, ESRF, CTH, DTU, CNRS, SOLEIL, CSIC
NATURE	Report
AUTHORS	C. Herrera, F. Cadiou, Q. Jacquet, N. Mozhzhukhina, A. Benayad, S. Lyonnard
CONTRIBUTORS	X. Liu, R. Fantin, D. Atkins, D. Blanchard, F. Capone, A. Naylor, P. Norby, L. Pérez Ramírez, A. Ponrouch, J. P. Rueff, S. Clark (WP7), E. Flores (WP7), J. Flowers (WP10), S. Fuchs (WP10)
CONTRACTUAL DEADLINE	30/11/2023
DELIVERY DATE TO EC	30/11/2023
DISSEMINATION LEVEL (PU/CO)	PU

ACKNOWLEDGMENT

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 957189. The project is part of BATTERY 2030+, the large-scale European research initiative for inventing the sustainable batteries of the future.

ABSTRACT

The BIG-MAP European project aims at revolutionizing next generation battery innovation, notably by greatly accelerating R&D capabilities. To advance this goal, BIG-MAP focuses on creating a database composed of modelling, electrochemical and characterization data collected in standard conditions using state-of-the-art experimental methods. This database will be used to train artificial intelligence models designed to accelerate materials discovery. Central to this database is the work of Work Package 5 (WP5), a cornerstone dedicated to advanced materials characterization. WP5 main goal is to design and build a European experimental platform performing multi-modal characterizations using standardized cells, protocols, data collection, data treatment and data analysis.

The large datasets produced in WP5 require robust methodologies that streamline data analysis while ensuring consistency and reliability. Data analysis standardization and automation involves the development of common protocols, formats and tools that enable seamless data interpretation across diverse experiments and instruments. Additionally, this will ensure that the generated data in BIG-MAP forms a robust and reliable foundation for further advances. In order to facilitate these critical functions, the standardization and automation of experimental data and observables emerge as an imperative.

The primary objective of this report is to present the data analysis conducted within WP5, focusing on the intricate landscape of data analysis. In WP5, data is produced in research laboratories and large-scale facilities, resulting in sizeable and varied datasets often requiring relatively complex post-processing and in-depth analysis. Here, we have selected specific techniques to demonstrate the need for standardization and automation based on real use cases. In line with this methodology, we developed protocols for data analysis standardization for XRD, XPS and Raman/FTIR techniques following specific schemes (i.e., discussions among experts and/or open round-robin analysis exercises for experts and non-experts). We performed automation for XRD, Raman, X-ray tomography segmentation and μXRD imaging datasets with procedures developed in WP5, such as data handling and analysis workflows, and novel analysis tools using machine learning approaches, as well as using more general automated analysis tools developed in collaboration with WP9. All of this was done using data formats defined jointly with WP9.

TABLE OF CONTENTS

1 Introduction

In this section, we provide a contextual overview of this deliverable. We present the organization and goals of WP5, introduce the objectives of the deliverable and describe the data analysis approaches.

1.1 Organization of WP5 and goal of the deliverable

WP5 centres on characterizing materials and battery cells through a combination of laboratory and large-scale facilities (LSFs) experimental methods. The aim is to establish the foundation for a European multi-modal experimental platform using standardized cells, protocols, metadata, data collection, data treatment and data analysis.

WP5 is organized in four tasks and five deliverables. Task 5.1 built an experimental matrix, published in deliverable D5.1, describing partners' capabilities and characterization techniques. The matrix-guided experiment selection for experimental workflows. Task 5.2 defined experiment categories, organized concurrent multi-site experiments, and established methodologies and proof-of-concept for a European multi-modal platform. Deliverable D5.5 documents the initial Tier 1 experimental workflow. The forthcoming deliverable D5.2 will expound on the proof-of-concept workflow and its future implications. Task 5.3 designed, prepared and executed Tier 2 experiments on a selected chemistry, demonstrating the effort on coordinating multi-technique and multi-scale experiments. Deliverable D5.3 showcases this experimental workflow. Lastly, task 5.4 is dedicated to the data analysis of experimental results derived from the aggregate of BIG-MAP experiments carried out within WP5. It seeks to standardize and automate data analysis across single and multi-techniques approaches, to provide uniform data and observables to the consortium. It includes the use of automated analysis modules developed in collaboration with WP9, the development of a novel data analysis tool for selected chemistry, and the resolution of big data challenges. The present report provides an in-depth description of task 5.4.

1.2 Overview of Task 5.4

Data generated in WP5 is complex and diverse in terms of structure, scale and origin, resulting from various materials, experimental workflows and scientific queries. To effectively analyze this data, we emphasize standardization, automation and correlation as crucial components, ensuring readability, reproducibility, and interoperability.

Given the intricate nature of WP5's data, we focus on specific use cases to illustrate the implementation of these approaches. Figure 1 provides an overview of the data analysis work within WP5. We have chosen to highlight four specific use cases: *crystallographic study of LNO* including the data analysis and automation of XRD data by CEA, DTU and ILL, *vibrational spectroscopy* focused on Raman and FTIR analysis by CTH*, post-mortem analysis* on reproducibility using XPS lead by CEA, and *tomography segmentation* including machine learning approaches by ESRF and DTU. Each one of these use cases was driven by specific scientific objectives, involving the development of tailored data analysis methods, tools and algorithms. Guidelines for data analysis standardization and automation have been shaped based on the insights gained from our data analysis effort in the frame of these use cases.

Analysis of the data obtained during the experimental workflow (D5.3) has started, and we have just begun addressing multi-technique correlation. This last undertaking is a forthcoming challenge that will reach maturity in future projects.

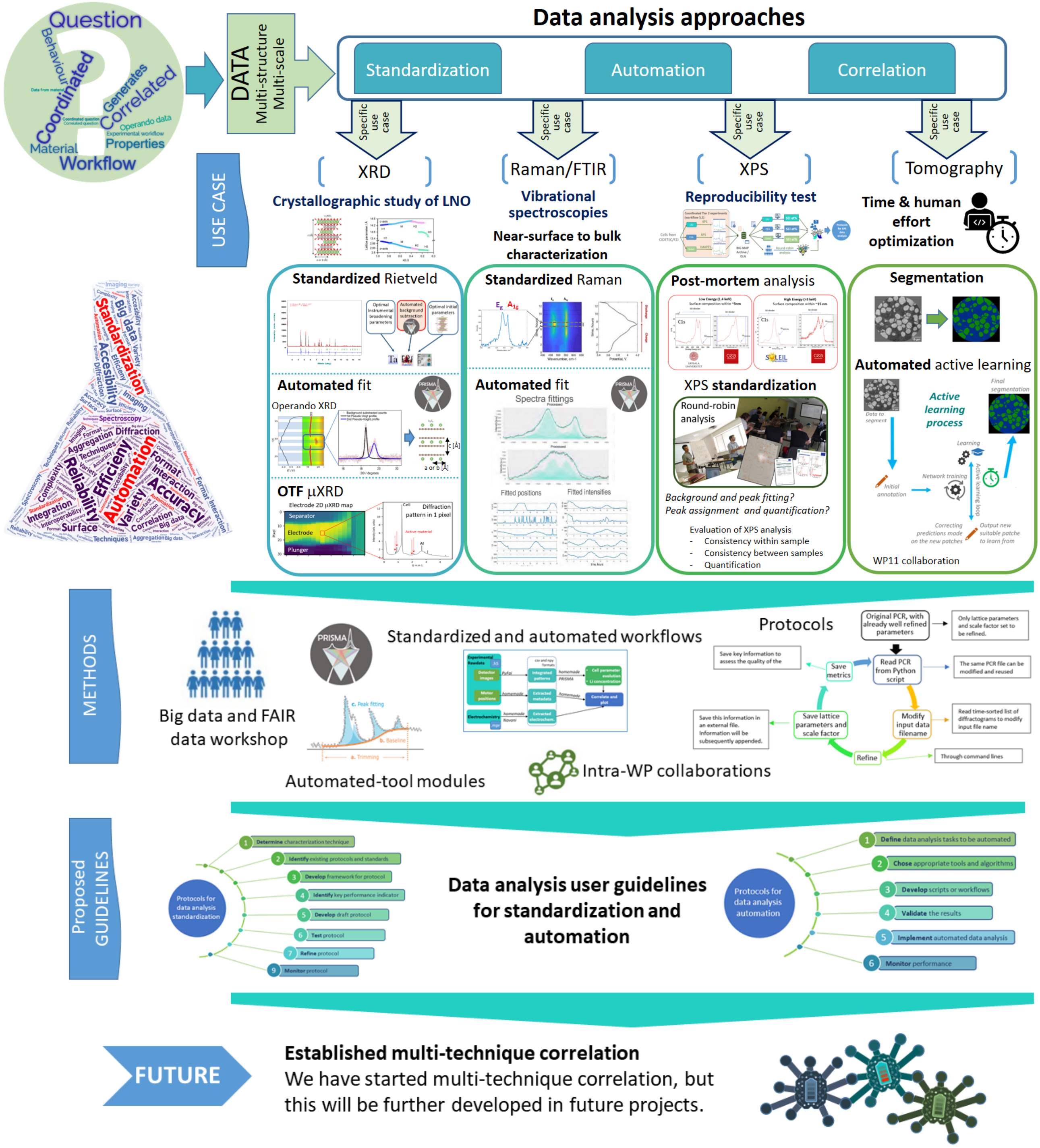

Question
Behaviour
Coordinated
Generates
Correlated
Operando data
Material
Properties
Workflow
DATA
Multi-structure
Multi-scale
Data analysis approaches
Standardization
Automation
Correlation
Specific use case
USE CASE
XRD
Crystallographic study of LNO
Raman/FTIR
Vibrational spectroscopies
Near-surface to bulk characterization
XPS
Reproducibility test
Tomography
Time & human effort optimization
Standardized Rietveld
Automated fit
Operando XRD
OTF µXRD
Electrode 2D µXRD map
Separator
Electrode
Plunger
Diffraction pattern in 1 pixel
Standardized Raman
Automated fit
Spectra fittings
Fitted positions
Fitted intensities
Post-mortem analysis
XPS standardization
Round-robin analysis
Background and peak fitting?
Peak assignment and quantification?
Evaluation of XPS analysis
- Consistency within sample
- Consistency between samples
- Quantification
Segmentation
Automated active learning
Active learning process
Initial annotation
Network training
Correcting predictions made on the new patches
Output new suitable patche to learn from
WP11 collaboration
Standardization
Accessibility
Diffraction
Spectroscopy
Efficiency
Automation
Reliability
Accuracy
Variety
Integration
Aggregation
Surface
Correlation
Big data
Techniques
METHODS
Big data and FAIR data workshop
PRISMA
Automated-tool modules
c. Peak fitting
b. Baseline
a. Trimming
Standardized and automated workflows
Protocols
Original PCR, with already well refined parameters
Only lattice parameters and scale factor set to be refined.
Read PCR from Python script
The same PCR file can be modified and reused
Modify input data filename
Read time-sorted list of diffractograms to modify input file name
Refine
Through command lines
Save lattice parameters and scale factor
Save this information in an external file. Information will be subsequently appended.
Save metrics
Save key information to assess the quality of the
Intra-WP collaborations
Proposed GUIDELINES
Protocols for data analysis standardization
1 Determine characterization technique
2 Identify existing protocols and standards
3 Develop framework for protocol
4 Identify key performance indicator
5 Develop draft protocol
6 Test protocol
7 Refine protocol
9 Monitor protocol
Data analysis user guidelines for standardization and automation
Protocols for data analysis automation
1 Define data analysis tasks to be automated
2 Chose appropriate tools and algorithms
3 Develop scripts or workflows
4 Validate the results
5 Implement automated data analysis
6 Monitor performance
FUTURE
Established multi-technique correlation
We have started multi-technique correlation, but this will be further developed in future projects.

Figure 1. Scheme for the data analysis work within WP5.

1.3 Data analysis approaches

Data analysis standardization

Reproducibility is a major principle of the scientific method [1, 2]. The results obtained after analyzing data from an experiment should be reproducible, i.e., we should be able to duplicate the results using the same materials (data) and procedures (methods) as were originally used [3]. In order to achieve this, we need to standardize not only experiments but also data analysis. Task 5.4 aims at standardizing protocols and workflows for the data analysis for individual techniques. This also implies standardizing the data format of all techniques, following common format defined in WP9 and ontologies defined in WP7 and thus enabling all users to access, understand and process specific data.

We have identified relevant techniques that demonstrate the necessity for data analysis standardization. Here, we elaborate on XRD, XPS and Raman/FTIR.

Data analysis automation

Automation is pivotal in data analysis, closely linked to standardization and big data. With the increase in data volume from various experiments (e.g., operando, tomography), automating data analysis is more crucial than ever. Manual processing is time consuming and prone to errors in every step in the process. Automation allows for fast processing and ensures accuracy and consistency in data analysis. Using programming languages such as Python simplifies the process. Scripts can handle tasks on data files automatically, even identifying patterns for data-driven decisions.

In this report, we focus on the data analysis automation of operando XRD, Raman and μXRD imaging data, and tomography segmentation through machine learning approaches. The need for automated tools to perform standardized spectroscopy data motivated the collaboration between WP5 and WP9. This resulted in the development of the **PRISMA**[1] software, an automated analysis module that can handle large amounts of spectral data in an efficient and reproducible way.

Data analysis correlation

Characterization techniques cover different chemical/structural properties and length/time scales within battery cell processes. Correlating complementary techniques is fundamental to understanding these mechanisms. Data analysis can also benefit from simulations and modeling, driving collaborative efforts with WP2 and WP3. Within this report, we present a successful collaboration using both experimental FTIR data and molecular dynamics (MD) simulations.

2 Use case: Crystallographic study of LNO

The *crystallographic study of LNO* use case was selected to showcase the implementation of the data analysis approaches. We include three characterization techniques in operando mode, XRD, μXRD imaging and NPD.

[1] PRISMA: a robust and intuitive tool for high-throughput processing of chemical spectra. PRISMA is an open source and user-friendly software to visualize and analyze spectra from in situ and operando experiments. It is based on Python for a reproducible, high-throughput processing of multiple patterns in an automatable way. The main features are the implementation of methods for trimming, baseline correction and peak fitting [4]. The code and a tutorial are available through the BIG-MAP App Store, https://big-map.github.io/big-map-registry/apps/prisma.html.

2.1 The scientific motivation

LNO ($LiNiO_2$) undergoes crystal structure changes during cycling, starting from a hexagonal phase (H1) in its pristine state, followed by transitions to monoclinic (M), then a second (H2) and a third (H3) hexagonal phase upon charge and delithiation of the material [5-7]. These phase transitions can result in multi-phase regions where more than one phase coexist. An operando approach is essential to characterize the crystal structure evolution and phase transitions during cycling. This is particularly significant as the behavior in the charged state significantly impacts materials performance with delithiation causing substantial volume shrinkage, primarily attributed to the H2/H3 transition (which alone accounts for a 4-5% volume change [8, 9]). The coexistence of phases with different volumes leads to interfacial strain, resulting in cracks along the primary and secondary particles [7]. The primary objective of XRD experiments in this case is to identify phases and estimate lattice parameters for LNO, specifically focusing on investigating lithiation and LNO phase transitions during battery cycling.

In this section, we present XRD data analysis standardization. Notably, our work introduces novel aspects as we demonstrate how traditional analysis software programs can be coupled with newer ones - such as the BIG-MAP app PRISMA - to automate background subtraction and perform an initial identification of phases during battery cycling. Furthermore, we illustrate methods to automate the refinement of multi-phase data and μXRD imaging in operando.

2.2 Implementation of data analysis standardization

2.2.1 Data analysis of powder diffraction

Neutron and X-ray powder diffraction (NPD, XRD) are complementary techniques to study the crystalline structure of a sample, i.e., the position of the atoms in the cell and therefore lattice parameters (see Annex 9.1). NPD and XRD data are represented in diffractions patterns, with intensity vs. scattering angle (2θ) as shown Figure 2.

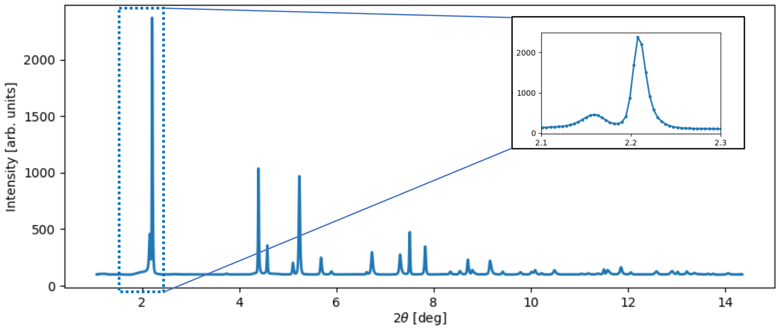


Figure 2. LNO diffractogram obtained at the ESRF beamline ID31. The diffractogram shows the scattered pattern of the X-rays in scattering angle 2θ vs. intensity. The insert zooms-in on one of the peaks.

The position and intensity of each peak in the diffractogram depends on various structural and experimental information: (1) crystal structure (units cell parameters, atomic parameters, etc.), (2) morphology (strain, stress, preferred orientation, etc.), and (3) instrumental information (incident wavelength, geometry configuration, sample alignment, etc.). Moreover, the measured diffraction pattern is a sum of the diffraction patterns produced by each phase (different type of materials) in a sample.

Data analysis is performed by minimizing the difference between the experimental data and reference patterns from databases, by changing structural/morphological/experimental parameters. The primary method for the full pattern fit is the Rietveld fit [10] used for refining crystal structures, lattice parameters, and reflections for all phases in the sample. This employs a non-linear least square approach to minimize a weighted residual function. A high level of quantitative phase analysis accuracy is reached by using the entire diffractogram.

Refinement software

Within WP5 we primarily use the open source software FullProf [11] and the commercial software Topas [12]. Both software packages can be used through a graphical user interface (GUI) and are scriptable. Topas input files (.INP) can be edited using JEdit, while FullProf input files (.PCR) can be modified with any text editor. Moreover, FullProf is executable through command lines, making it suitable for integration into customized Python scripts.

Detailed guidelines to perform general refinement have been described by the International Union of Crystallography Commission on Powder Diffraction [13]. Based on these recommendations, together with specific Topas[2] and FullProf[3] best practices, we developed specific protocols to refine BIG-MAP LNO data that can be extrapolated further to other chemistries.

A priori information for refinement

Before refining, the following sample and experimental condition information are required:

- diffractometer parameters and experimental conditions (incident wavelength, zero-shift, and instrument resolution) are obtained by measuring a reference sample under the same conditions as the rest of the experiment.
- Chemical composition and crystal structure information (space group, approximate unit cell parameters and atomic positions) are extracted from crystallographic datasets available online.

2.2.2 Defining protocols for XRD data analysis standardization

Methods

Figure 3 portrays the systematic approach we have adopted to develop protocols for XRD data analysis on LNO cycling data. During Tier 1 experiments, XRD experiments were carried out at MAX IV by DTU and at the ESRF by CEA. The produced datasets were centrally stored within the BIG-MAP archive[4] and metadata in the BIG-MAP Notebook[5], ensuring accessibility for all consortium partners. DTU, CEA and ILL performed independent and parallel data analysis, producing a report documenting the details to reproduce the refinement process. Annex 9.2 presents the report template. The team involved in this data analysis effort comprised partners with diverse experience in refining powder neutron and X-ray diffraction data, each contributing with their expertise in the data refinement.

Following the parallel data refinement process, we deliberated on refinement strategies, analyzed outcomes, and ultimately arrived at a consensus on common best practices for refining XRD data in the context of BIG-MAP batteries.

[2] https://topas.webspace.durham.ac.uk/

[3] https://www.ill.eu/sites/fullprof/

[4] https://archive.big-map.eu/

[5] http://big-map-notebook.u-picardie.fr

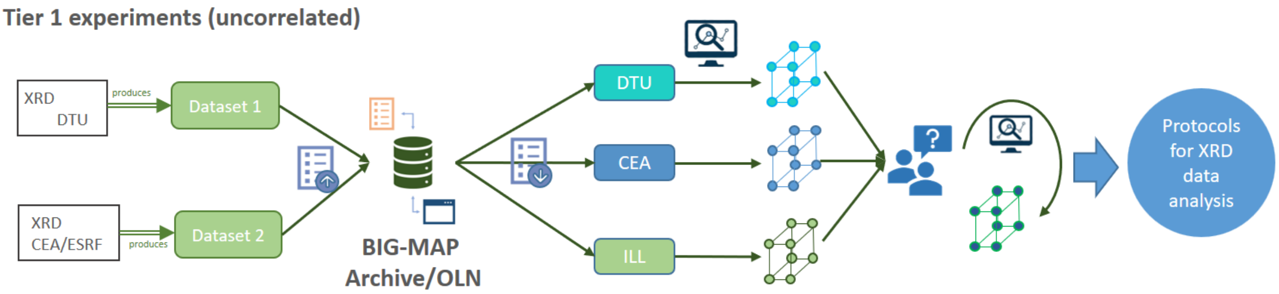


Figure 3. Illustration of the process used for defining protocols to standardize the XRD data analysis for the selected chemistry. From left to right, two institutions conduct XRD measurements, generating datasets that are uploaded to the BIG-MAP archive and metadata uploaded to the BIG-MAP Notebook, making them available to the entire consortium. Partners download one (or multiple) datasets and conduct independent data analysis. The results are compared, and interactive discussions are established to converge on protocols for XRD data analysis.

Protocols for XRD data analysis

The XRD data analysis process can be approached using different methods, depending on the data origin and user preferences. We report data obtained at the DanMAX beamline (MAX IV), for which standards and empty cell measurements are required; and at the ID31 beamline (ESRF), where, typically, standards are measured by the beamline scientists and empty cell measurements were not necessary since there were no amorphous background contribution. Approaches to refine these data share the same strategy. However, DanMAX data had an additional process in order to account for the background contribution. XRD peaks from Al, Cu and glassy carbon background were refined first, and refined parameters (lattice parameters, scale factors and thermal displacement parameters of Al and Cu atoms) were fixed for the following part.

Figure 4 illustrates the guidelines we have followed for refining ex situ XRD data. We emphasize three initial aspects: (1) accurate instrument calibration and zero-point determination, (2) close matching of initial lattice parameters values to the real ones, and (3) proper consideration of the background signal. The first two aspects are standard procedures, while background subtraction can be optimized by integrating traditional and specific refinement software packages with more versatile versions. Additionally, refinement has to be done in steps since some parameters have a greater impact on the overall shape and larger scales (i.e., scale factor, background) than others. Also, during refinement, after each iteration cycle, the refinement must be assessed equally at both low and high scattering angles, and metrics should be also followed.

Important considerations before refinement

- The different phases in the sample must be identifiable. For LNO: phase transitions from H1, M, H2 to H3.
- To reach convergence during refinement, requisite a priori information is needed (e.g., lattice parameters of the different phases).

This information is based on relevant literature, including references [5-7].

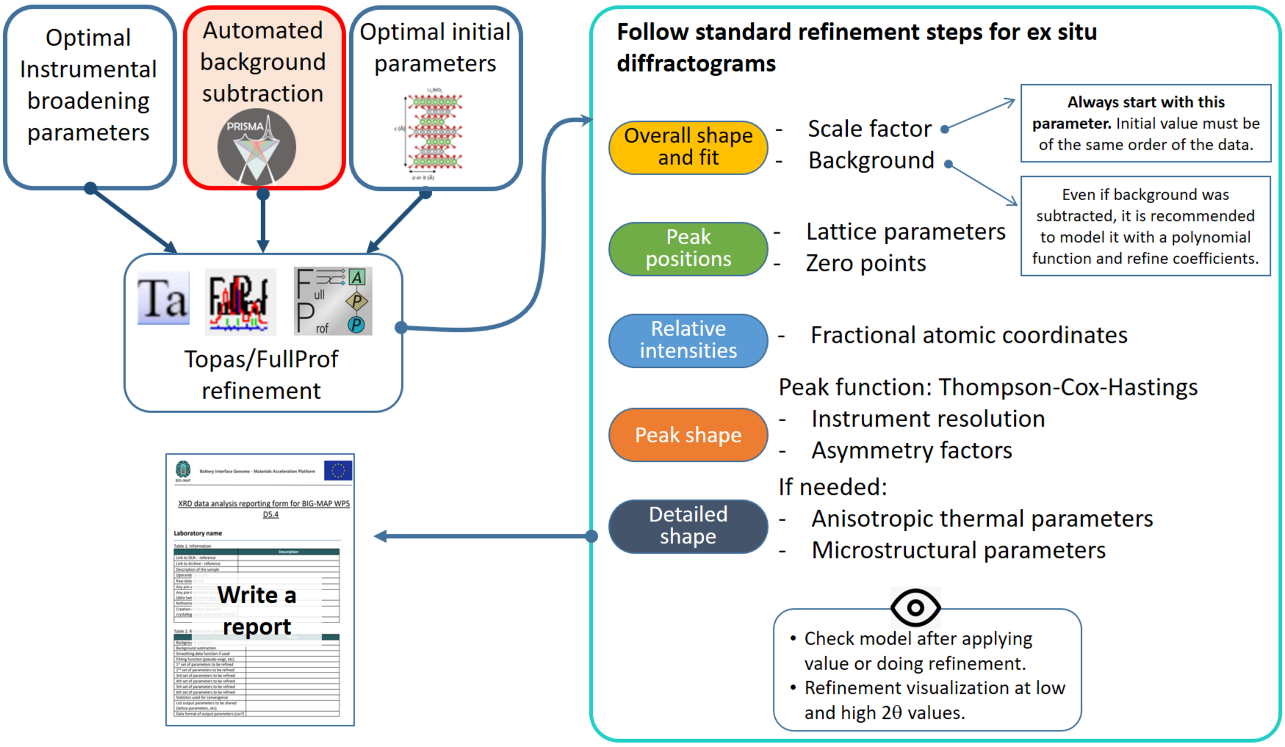

Optimal Instrumental broadening parameters
Automated background subtraction
PRISMA
Optimal initial parameters
Topas/FullProf refinement
Ta
Full Prof
Write a report
XRD data analysis reporting form for BIG-MAP WP5 D5.4
Laboratory name
Follow standard refinement steps for ex situ diffractograms
Overall shape and fit
- Scale factor
- Background
Always start with this parameter. Initial value must be of the same order of the data.
Even if background was subtracted, it is recommended to model it with a polynomial function and refine coefficients.
Peak positions
- Lattice parameters
- Zero points
Relative intensities
- Fractional atomic coordinates
Peak function: Thompson-Cox-Hastings
Peak shape
- Instrument resolution
- Asymmetry factors
If needed:
Detailed shape
- Anisotropic thermal parameters
- Microstructural parameters
• Check model after applying value or doing refinement.
• Refinement visualization at low and high 2θ values.

Figure 4. Schematics for the XRD data analysis process.

Background subtraction

In our workflow, we have integrated PRISMA for background subtraction. For this, we have chosen the Asymmetric Least Squares (AsLS) method developed by Eilers and Boelens [14, 15]. The AsLS method offers an excellent approach for baseline estimation. The method finds a curve that is both smooth and faithful to the spectrum while asymmetrically penalizing positive residuals, where the analytical peaks are found. The optimal baseline signal is determined through an iteratively minimization of a penalized least squares function. Within PRISMA, two parameters are tuned: penalty p and smoothness λ. Penalty p, linked to the weights of the data, emphasizes the relative importance of data points at the spectrum's base. Smoothness λ controls the smoothness of the baseline curve, for which larger λ values produce flatter curves (we refer the reader to the PRISMA publication for more information). Performing the background subtraction with PRISMA allows for fast automation during operando experiments (see Section 2.3.1.1).

This approach is optimal for LNO patterns, since there is no need to manually identify regions for interpolation. Many parameters are sensitive to background (e.g., scale factor, occupancy parameters, thermal parameters), thus optimizing background subtraction is key.

Reporting

For each data analysis process, a report detailing all steps carried out is mandatory. This provides the basis for documenting the entire analytical process. Through clear articulation of methodologies, data sources, assumptions and conclusions, the report provides a transparent record for future reference and replication.

2.3 Implementation of data analysis automation

2.3.1 Operando XRD

Operando XRD experiments produce thousands of diffractograms throughout battery cycling, each of them consisting of several peaks that represent the electrode structure during lithiation and de-lithiation processes. Automating data analysis is then time-efficient besides ensuring accuracy. It is crucial to adhere to specific protocols to handle and treat the data appropriately, as consecutive datasets are correlated.

In this context, we describe two approaches towards the automation of XRD data analysis. The first one uses PRISMA to automate the analysis of individual peaks. The second one involves sequential Rietveld refinement applied to the entire diffractograms, enabling the refinement of crystal structures and phase quantification in a sequential manner.

2.3.1.1 Automated data analysis with PRISMA for single peaks

While the Rietveld refinement method is well-suited for fitting entire diffractograms, the PRISMA software is specifically suited for fitting individual peaks.

PRISMA was used to fit the (00ℓ) reflections of LNO, as well as the electrode-independent background from operando XRD data obtained at DTU. Specifically, the 003 reflection was used. Its shift in position during cycling corresponds to a change in the lattice parameter *c*, which is an indicator of phase changes in the electrode, notably due to changes in lithium concentrations. After visual inspection, a specific approach was adapted. For the background, we follow the directions described in Section 2.2.2. For the peaks, two pseudo-Voigt profiles were applied, one at low (18.4-18.7 degrees) and high (19.5-19.7 degrees) angles, corresponding to the H1/M/H2 and H3 phase regions, respectively. Figure 5 displays this fit to a single diffractogram.

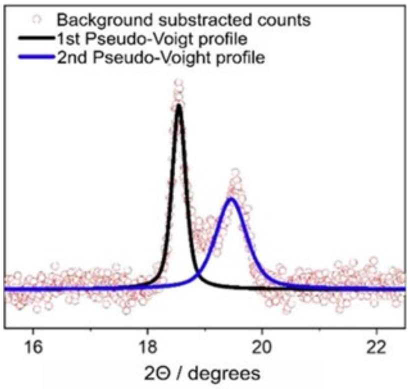


Figure 5. Extracted from Figure 6 in Flores et al. [4]. Zoom to the 003 reflection of a single XRD diffractogram of LNO obtained during operando experiments in biphasic region. Background-subtracted data is shown in red circles. The black and blue lines are fits to the H2 and H3 phases, respectively.

This approach is purely mathematical and does not require any knowledge of contiguous (temporal) data nor the physical properties of the sample. Fit convergence is quickly achieved using the same initial assumptions for all data, since the positions of the two components do not vary significantly. This analysis is presented in the PRISMA paper [4].

The processing of the XRD operando data was done using the PRISMA API in Python. The raw data, formatted in *xy* format, is compatible with PRISMA's reading capabilities. PRISMA processes several thousand patterns in just a few minutes. In this case study, the main objective of using PRISMA was to perform a rapid analysis focused on a specific scattering angle range within the diffractogram, which exclusively includes the targeted peak of interest. The evolution of the *c* lattice parameter during cycling allowed identification of the different phases.

2.3.1.2 Automation of sequential Rietveld refinement

Semi-automated Rietveld refinement of operando XRD data was done to obtain the evolution of all lattice parameters during cycling. The multi-phase nature of the material during cycling implies an inherent system complexity, and sequential refinement cannot be fully automated, i.e., some manual intervention is still required to ensure the accuracy of the refinement results.

From ex situ to operando sequential refinement

In Section 2.2 we discussed about the need for standardizing XRD data analysis and presented a scheme for standardized analysis of ex situ XRD data. Operando sequential refinement builds on these protocols. However, as stated, sequential Rietveld refinement of multiphase samples cannot be completely automated. Sequential refinement strategies rely on the smooth transition of patterns along the evolution axis, which, in our case, is time/cycling. During cycling, LNO phases shift from H1 to M to H2 to H3, with biphasic regions appearing, often resulting in abrupt pattern changes. In consequence, sequential Rietveld refinement involves a semi-automatic process, as manual input is required to identify when the phase transitions occur. Indeed, the refined parameters from one-time step are used as starting parameters for the subsequent diffractogram refinement. The most reliable way to perform the phase identification is by fitting a single peak through all diffractograms obtained during cycling, as discussed in Section 2.3.1.1. We can thus identify the scan numbers where the phase transitions occur and/or set limits for the lattice parameter sizes for each phase.

During the sequential refinement, only two sets of parameters for each phase are refined: **lattice parameters** and **scale factors**. All other parameters undergo zero or negligible change over the cycling and can thus be set as *fixed*. As discussed, instrumental (broadening) resolution parameters are previously refined against a standard reference. Accurate starting crystallographic structure parameters for each phase are also needed and must be determined from an initial XRD pattern.

The implementation of this strategy is software dependent. In any scenario, two constraints are applicable to all software:

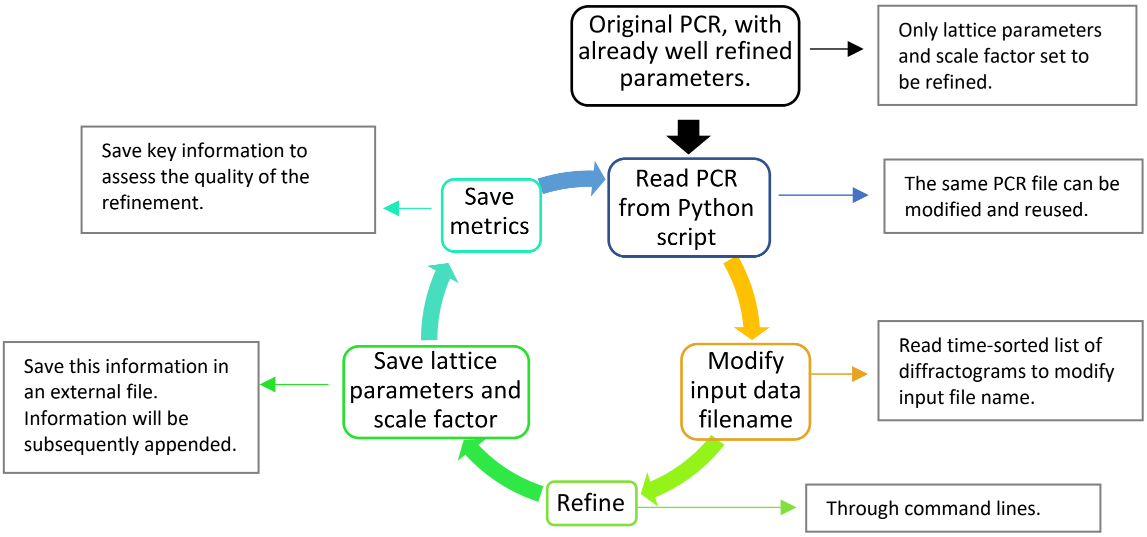
- The list of diffractograms provided to the software must be chronologically sorted.
- The identification of phase transitions and the limiting values of the lattice parameters are needed.

➢ In the Topas software, the process of sequential refinement is straightforward due to the existence of specific macros developed for this purpose. Additionally, Topas can handle conditional statements and parameterized variables, which is essential for handling phase transitions. With Topas:
  - A unique input INP file is used. The file list with the experimental data is provided.
  - The crystallographic structure information for all phases (H1, M, H2 and H3) is written.
  - Constraints for lattice parameters are imposed to avoid overfitting. For individual phases, scale factors are fixed to zero if certain lattice parameters reach a limit value, e.g.,

```
‘update =If(Or(LNO_1_a==2.86, LNO_1_c==14.27), 1e-10,Val );
        min = 1e-10;'1.37829701e-05_2.58493653e-07’
```

- During the sequential refinement, only the lattice parameters and scale factors for the respective phases are refined.

➢ In the FullProf software, the execution of sequential refinement can be obtained by developing specific Python scripts and using FullProf via command lines. Figure 6 illustrates the sequential refinement process. The steps are as follows:
    - Prepare an individual input file, PCR, for each phase and multi-phase configuration. For the first scan of each phase, refine parameters following protocols described in Section 2.2.2. Then, set to *fixed* all parameters but lattice parameters and scale factors.
    - Create a list of diffractograms to be refined sorted by timestamp, which will be provided to (or created by) a Python script. These will be used as *filename* variables to be modified in the PCR files at each iteration.
    - In the Python script, read the starting PCR file. Modify the data file name of the diffractogram to be refined according to the list. Refine and save the ensemble lattice parameters, scale factors and file name values in an external file.
    - For each subsequent diffractogram, modify the data file name and use the refined lattice parameters from the previous iteration as the new initial assumption then perform the refinement. Repeat this process iteratively until all diffractograms are processed.
    - This entire sequence is repeated for each phase or multi-phase configuration.
    - For multi-phase regions, a critical assessment of the phase fraction is essential to verify the effectiveness of the refinement process. Particular attention must be given to phases within multi-phase regions that show low phase fractions.



Original PCR, with already well refined parameters.
Only lattice parameters and scale factor set to be refined.
Save key information to assess the quality of the refinement.
Save metrics
Read PCR from Python script
The same PCR file can be modified and reused.
Save this information in an external file. Information will be subsequently appended.
Save lattice parameters and scale factor
Modify input data filename
Read time-sorted list of diffractograms to modify input file name.
Refine
Through command lines.

Figure 6. FullProf sequential refinement process.

Important considerations

- Carefully refine the reference source and/or the initial diffractogram to obtain the most accurate instrumental values.
- Starting parameters for refinement are always necessary.
- Background should be accounted for accordingly, following methods described in Section 2.2.
- Assess the quality of the refinement by storing metrics.

2.3.2 Operando synchrotron µXRD imaging: a workflow for (semi) on-the-fly analysis

2.3.2.1 Why do we need automatized data analysis?

Reaction kinetics in batteries are heterogeneous at multiple length scales (particle, electrode, cell). Spatially resolving these heterogeneities is necessary to better understand how batteries work, and crucial to calibrate battery modelling. Heterogeneities in the depth of the electrodes have been measured in WP5 using operando synchrotron µXRD imaging at the ESRF. A micron-sized beam is scanned across an electrode producing a 2D map of the electrode in which every pixel contains a diffraction pattern. Typical pixel sizes are 3 µm in the electrode depth and 100 µm in plane. Each map takes approximately 5 min to acquire. Our operando experiment lasted 3 days and over 10^5 detector images were produced. Clearly, an automated way of analyzing this dataset is necessary. Figure 7 illustrates the generated data complexity.

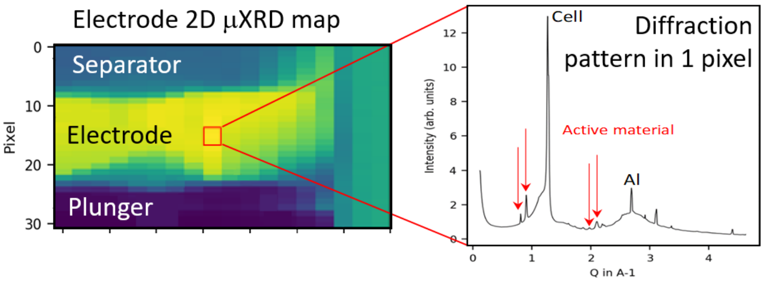


Figure 7. Example of the generated data from our µXRD imaging experiment. Left side: the 2D map for one of the active material peaks. Right side: diffractogram at a position in the electrode.

2.3.2.2 Workflow structure

We have developed a workflow based on several Jupyter Notebooks built using in-house and existing modules enabling a nearly on-the-fly analysis of operando µXRD imaging. It takes about 4 hours to set up the data treatment workflow for a given experiment. The structure of the workflow is shown Figure 8, and it comprises: (1) detector image integration using PyFAI[6], (2) analysis of the diffraction patterns using PRISMA to obtain cell parameters and/or phase fractions, (3) extraction of metadata (motor positions, time, beam flux and energy), (4) import of electrochemical data (using Navani[7]), (5) direct correlation of structural parameters, metadata, and electrochemical data, (6) and visualization. The workflow is used on a regular basis on beamlines BM32, BM02, ID13 and ID31 at the ESRF.

All processed data are saved in csv (human readable) and npy (machine readable) formats. Metadata of the analyzed, as well as that of electrochemical data, are saved in json files. This allows for experimental and data analysis reproducibility. Future development of the workflow includes the merging of data and metadata into HDF5 files, replacing the npy and json files, and the addition of two techniques to the workflow, Small Angle X-ray Scattering imaging analysis and scattering tomography. The workflow will be released as open source on GitHub.

[6] https://pyfai.readthedocs.io/en/v2023.1
[7] https://github.com/the-grey-group

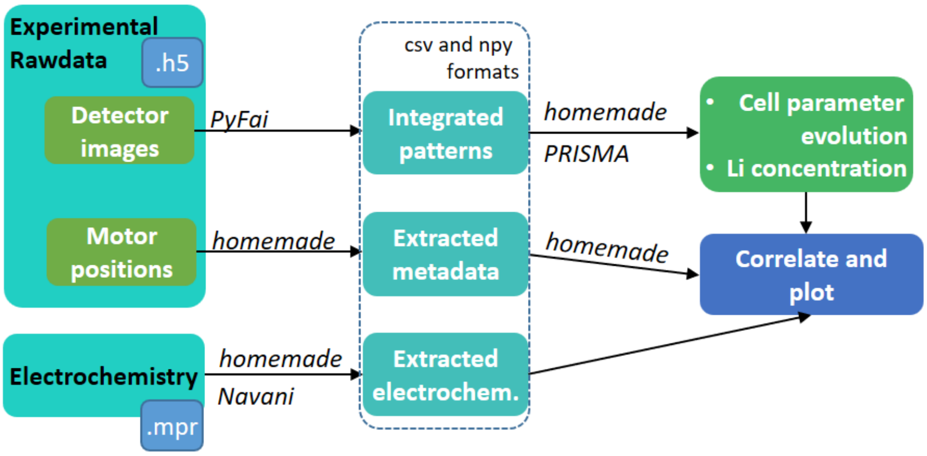


Figure 8. Schematics of the automated workflow for analyzing operando µXRD imaging.

2.3.3 Analysis automation for NPD data at the D2B instrument at the ILL

In the context of the on-the-fly visualization of NPD data, as described in Deliverable D5.3 (Section 3.5), we have developed algorithms for automated data reduction and statistical analysis of diffractograms acquired at the D2B instrument at the ILL.

The algorithms were constructed in two stages. Initially, fundamental algorithms for data reduction of individual measurement and pairwise diffractogram comparisons were written in a single file, `D2Bdatareduction.py`. Subsequently, a second script, `automated_datacal.py`, which calls the former script, was created to automate the data reduction process for all numors (raw data NeXus files) acquired during an experiment, along with a statistical analysis of the sequentially obtained diffractograms producing a PDF report.

2.3.3.1 Automated data reduction

We have outlined the specifics of data reduction and visualization for a single D2B measurement in Deliverable D5.3, Section 3.5.2.2. To summarize, one D2B measurement generates ten files, each containing twenty-five data matrices. In the following, we describe the data reduction automation procedure for all files obtained in an experiment.

Automation relies on keywords in the numors' metadata. There are two essential keywords used for this automation:

- `experiment_identifier`: string variable which contains the experiment name provided by the user at the time of measurement. This parameter is used by the script to assign names to the resulting calibrated diffractograms, workspaces, as well as saved figures.
- `sequence_id`: float-point variable that represents the time at which the measurement sequence was initiated. Files sharing the same `sequence_id` value were generated during the same measurement session, Detector Scan mode, requiring to be processed together.

The `automated_datacal.py` script can be executed to apply data reduction to all numors in the `rawdata` folder by running: `python automated_data.py`. Calibrated 1D diffractograms in NeXus and text formats will be saved in the autogenerated `Calibrated1D-nxs` and `Calibarted1D-txt` folders, respectively.

To process the data, numors are grouped according to their `sequence_id` values and processed together. Calibrated diffractograms are saved using the `experiment_identifier` variable as file name.

2.3.3.2 Automated data analysis

The primary purpose of conducting the initial data analysis on the diffractograms from operando experiments is to detect any abrupt changes in the patterns, which may indicate phase transitions. The developed scripts were designed to provide an initial and quick overview of the differences between two diffractograms.

The `Comparison_spectra()` function performs a statistical comparison of two numors. It accepts two string variables, which should be the names of two 1D NeXuS processed diffractograms located in the `Calibrated1D-nxs` folder. This function computes the subtraction of the two diffractograms and the division of one by the other. From the subtracted plot, it estimated the Root Mean Square Error (RMSE)[8] and Pearson correlation coefficient[9]. Additionally, we also compute the Kolmorogov-Smirnov statistics[10] between the two diffractograms. From the divided curve, statistical values are calculated including standard deviation, minimum, maximum, mean and median values. The function also generates a figure displaying three plots: an overplot, subtraction and division of the diffractograms (see Figure 9).

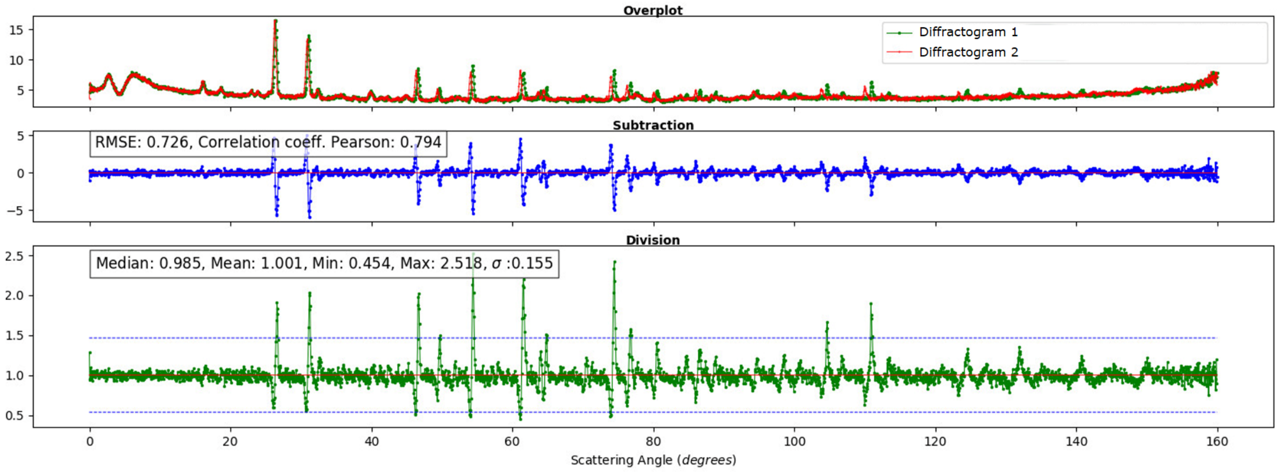


Figure 9. Example of graphical comparison between two diffractograms. Top panel: overplot of the two diffractograms. Middle panel: subtraction of the two diffractograms. RMSE and Pearson coefficient are displayed. Bottom panel: division of the two diffractograms. The median, mean, minimum, maximum and standard deviation of the curve are given. Red horizontal lines represent the mean values, and the two blue-dash horizontal lines the ±3σ_{dev}+μ_{mean} values.

By graphically and statistically visualizing the difference between two consecutive diffractograms, we can track changes in phases throughout cycling. This function can also be used to compare NeXus calibrated diffractograms, obtained from other instruments.

While the `Comparison_spectra()` function can be directly used to compare two specific diffractograms, it is also integrated into the automated data calibration process. When multiple

[8] The RMSE is an absolute measure of the difference between the two diffractograms and is calculated as the square root of the mean of the squared differences between the two diffractograms values. It quantifies how far, on average, the two y-axis from each diffractograms are.

[9] The Pearson correlation coefficient measures the strength and direction of the linear relation between the two diffractograms. Low values of RMSE (close to zero), as well as high Pearson coefficient (close to one), indicate that the diffractograms are similar.

[10] The Kolmorogov-Smirnov (KS) statistic quantifies the maximum vertical distance between the cumulative distribution functions of the two diffractograms being compared, i.e., the largest discrepancy between them. Small values (close to zero) indicate that the two diffractograms are similar in their distribution.

diffractograms require data calibration, the comparison is conducted between two consecutive measurements in time. The output is provided in the form of a PDF file containing figures and statistical specifications.

The scripts can be found on GitHub[11].

2.4 On-the-fly data analysis tool

We developed a data analysis tool employing a machine learning approach to perform on-the-fly phase identification in LNO electrodes during operando XRD experiments. The methodology for this work involves:

1. Collect operando XRD data
2. Refine the data and identify phases
3. Aggregate electrochemical data to associate lattice parameters with lithiation values (*x* in Li_xNiO_2)
4. Extract lattice parameters, corresponding phases and lithiation
5. Generate thousands of sets of phase-labeled lattice parameters, randomly generated within a specific range defined by the experimental data
6. Simulate diffraction patterns from the generated lattice parameters
7. Train convolutional neural network models on the simulated data
    a. Classifying phases (this work)
    b. Regressing *x* (future work)

2.4.1 Experimental data handling and simulated data

XRD data

We use X-ray diffraction data from the ESRF (beamline ID31) for our analysis. Operando XRD data was sequentially refined using FullProf, and phase identification was conducted. Aggregation of the electrochemical data was done by matching the closest timestamps in the datasets with custom scripts. This process enabled us to extract refined lattice parameters, phases, and lithiation for each scan in the operando experiment.

Generation of interpolated/randomized data

We generate lattice parameters by means of interpolation and randomization using the refined lattice parameter values. For each phase (and biphasic area), we extract the minimum and maximum *x* values. We then iterate N times, generating a random *x* value within these limits during each iteration. For each randomly generated *x* value, we find the nearest lower and higher *x* values, and extract the minimum and maximum values for each lattice parameters within this *x* range. These values were used as constraints for generating random lattice parameters.

Figure 10 illustrates an example of generated lattice parameters of the M phase for N=50 and N=500. Circles represent the refined data while green crosses correspond to the generated values. In theory, each *x* value should correspond to a unique set of lattice parameters. However, due to the rounding of floating-point numbers by different algorithms and the uncertainties in the refinement process, we observe some effects in the refined parameters in Figure 10. Typically, these values are averaged for better representing the data but here we have chosen to show all data. We selected our strategy for generating the lattice parameters to encompass a range of the typical values.

[11] https://github.com/cnherrera/automated-data-calibration-d2b-mantid

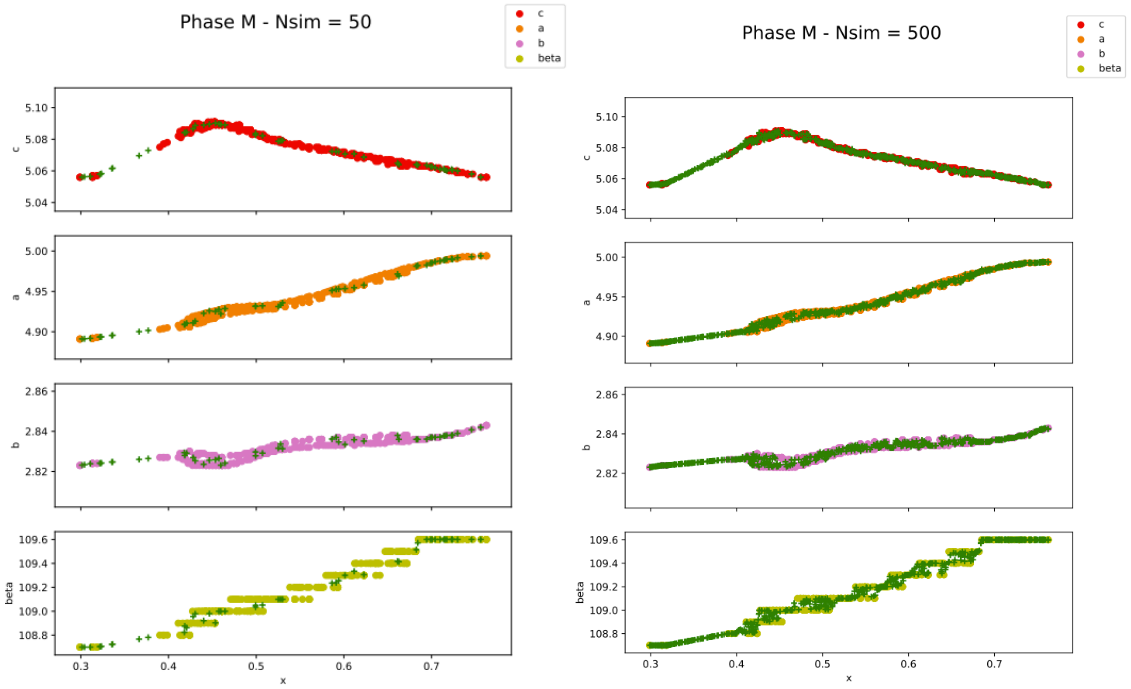


Figure 10. Example of generated lattice parameters from a limited number of refined ones. Circles correspond to refined values from the experimental data, while green crosses correspond to the generated parameters. Left and right panels display the generated values for 50 and 500 iterations, respectively.

Simulated diffractograms

We use FullProf in the command line version `fp2k` and executed through a Python script to simulate X-ray diffraction patterns from the generated lattice parameters grid. At each iteration, the input PCR file was modified by (over)writing the lattice parameter values and running the refinement. Due to the time-consuming nature of this process, parallel processing was applied to handle all phases simultaneously. The resulting simulated patterns for each phase were saved in individual h5 files.

2.4.2 Data preparation for machine learning model

Data needs to be appropriately pre-processed to make it suitable for machine learning models. Here we outline the steps we take to prepare for training and test:

1. Read the datasets (simulated diffractograms) from external files.
2. Extract a subarray within a fixed 2θ (q) range and resolution to match and standardize all data.
3. Fit and subtract the background using the AsLS method, and introduce scaled Poison noise.
4. Shuffle the data to ensure randomness.
5. Split dataset into train and test data.
6. Normalize patterns for consistent scaling.
7. Apply label encoding for classes.

Outputs are four variables needed for the machine learning model, X_train, X_test, y_train, y_test.

2.4.3 Convolutional Neural Network model

We have chosen a Convolutional Neural Network (CNN) for our data training, as convolutional layers are well suited for capturing local patterns like peaks or free-peaks regions and hierarchies. The model architecture comprises two 1D convolutional layers for feature extraction, two dense hidden layers to enhance learning capacity and an output dense layer for categorical predictions.

To enhance the robustness of the model and prevent overfitting, we introduced MaxPooling layers to downsample the spatial dimensions and incorporated dropout layers throughout the training process. We systematically evaluate the evolution of training and validation loss and accuracy. We employed confusion matrices to assess the performance of the model.

The entire model is implemented using TensorFlow/Keras. Details on the final architecture, including specific hyperparameters chosen, as well as on the simulated data, will be published and made available in the BIG-MAP GitHub.

3 Use case: Post-mortem analysis

The *Post-mortem analysis workflow* use case focuses on reproducibility tests from cell assembly to data analysis. Our objective is the **data analysis standardization** of X-ray photoelectron spectroscopy, a surface technique particularly useful to study the SEI chemical composition on graphite electrodes upon cycling.

3.1 The scientific motivation

X-ray photoelectron spectroscopy (XPS) is a surface technique that can be carried out at laboratories and synchrotron facilities. It is used to investigate the chemical composition of surfaces within the first 5 nm depth. XPS can probe the SEI composition of a battery cell electrode by revealing the chemical bonding of the SEI by-product's nature and their semi-quantitative atomic percentage (see Annex 9.1 for more details). There is considerable understanding of the SEI on graphite electrode using carbonate-based electrolyte mixed with $LiPF_6$ salt, thanks to XPS core level analyses of C 1s, O 1s, F 1s, P 2p and Li 1s evolution upon lithiation/de-lithiation [16]. However, data analysis carried out on the same spectra recorded by the same equipment and treated by different XPS experts using different software is not well established and not well documented.

We will focus here on the standardization of the XPS data analysis procedures.

3.2 XPS technique and data analysis

The XPS spectra represent the number of photoemitted electrons per second scanned over the total kinetic energy spectra. Survey mode spectra, obtained at low resolution, allows the identification of core peaks and to obtain an initial overview of the sample's chemical composition. Specific features such as asymmetric background emission can also be identified. High resolution spectra at individual core peaks are used for data analysis.

In WP5, there are three partners that conducted XPS experiments: CEA, Uppsala (UU) and SOLEIL (synchrotron). Table 1 provides an overview of the XPS instruments, including their specifications and energy source. Different source energies have different penetration depth capabilities, producing complementary information across the surface layers.

Various software packages, such as CasaXPS, XPSpeak, MultiPack, and AAnalyzer, can be employed for data analysis. In WP5, all three partners use the commercial CasaXPS software[12].

[12] http://www.casaxps.com/

Table 1. Overview of the instruments in WP5 that perform XPS/HAXPES measurements.

Partner	Instrument	Source	Energy	Probing depth	Charge Neutralization
CEA	QUANTES (PHI)	Al kα	1486.6 eV	~2-10 nm	Available
		Cr kα	5414.8 eV	~5-20 nm	
UU	PHI 5500	Al kα	1486.6 eV	~2-10 nm	Available
SOLEIL	Galaxies beamline Scienta EW4000	Synchrotron	2300 to 12000 eV	~ 10-50 nm	Not available

Data analysis is done by peak fitting. It is subjected to several approximations related to the background subtracting methods, the selected base line shape and the mathematical function used to fit the spectra. Despite the fact that this approach is widely used (but not standardized), the physical background of different steps used in the treatment of spectra can strongly affect the qualitative and quantitative interpretation of the collected results, and thus the ultimate understanding of the SEI composition.

Peak fitting is used to separate the intrinsic photoemission signal from different extrinsic photoemission related processes. This approach allows the identification of different constituent chemical states, facilitating comprehensive data interpretation. Prior to this process, a peak calibration is performed relatively to the Fermi level position of the sample. In general, the C 1s peak assigned to carbon contamination is used for this purpose. For a chemically heterogeneous surface, such as the SEI, it is usual to use the peak of the most representative chemical species measured. This process involves two main components: background and peak fitting, which are displayed in .

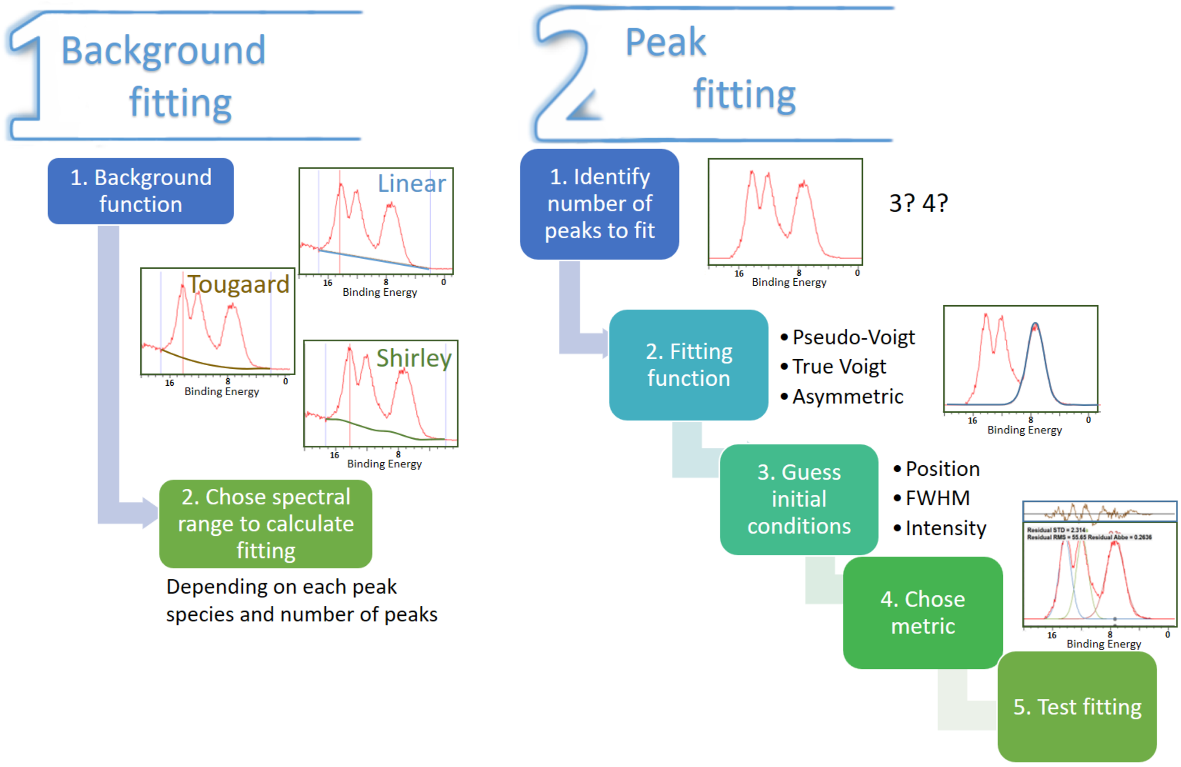


Figure 11. Process for XPS data analysis.

XPS background emission is sample-dependent, it is mainly due to inelastic scattered electrons within the sample. It is often modeled with linear, Shirley and Tougaard functions[13]. The spectral range choice (trimming) to perform background fitting is also critical.

For peak fitting, the first step is to determine the number of peaks, followed by the selection of a fitting function (e.g., pseudo-Voigt, true Voigt, or asymmetric function). Providing accurate initial values for, respectively, peak position, full-width-at-half-maximum (FWHM) and intensity is crucial for convergence. From a mathematical point of view, fit quality can be evaluated using metrics like root mean square (rms) or residual standard deviation (χ^2), whereas from a chemical aspect, coherence of chemical structure must be highlighted by cross-linking information from corresponding core level peaks (for example, based on reference study, the LiF species has F 1s and Li 1s peak position at 685 and 56 eV, respectively). Additionally, a coherence in the stoichiometry must be respected through semi-quantitative analyses. Expert XPS users may validate results by comparing the fitted peaks with other expected peaks in the data.

The results of the fitting model of each peak, position, peak and FWHM, give us the following information:

1. Position – binding energy: elemental and chemical state of the atoms through binding energy shift. The XPS is the most direct non-destructive technique that allows chemical bonding identification within 0.5 eV resolution for a modern lab-based spectrometer.
2. FWHM: information concerning the chemical state and the chemical and/or physical environment of the atom. Since the FWHM is the signature of photoexcited hole lifetime (Lorentzian distribution) and instrumental broadness (Gaussian distribution), the use of pseudo-Voigt function is recommended to describe covalence bonding related peak signature.
3. Peak intensity: amount of substance, i.e., quantitative information about the concentrations of the chemical states. Relative sensitivity factors (Scofield factors) are used to scale the measured peak areas so variations in the peak areas are representative of the amount of material in the sample surface. Nowadays relative sensitivity factors are available for Al and Mg X-ray based XPS. The increase of interest in lab-based HAXPES techniques is paving the way for building new relative sensitivity factors for high X-ray energy.

3.3 Implementation of data analysis standardization

In order to inquire and ensure standardization of the XPS data analysis among our partners, a post-mortem analysis in two stages was conducted. The first stage focused on the reproducibility of ex situ samples in WP5 laboratories, and it was developed during Tier 1 experiments. Preliminary results from this stage were inconclusive and thus a second stage was developed, integrated in the 5.3 experimental workflow based on Tier 2 experiments. The following sections detail this two-stages post-mortem analysis.

3.3.1 Post-mortem analysis 1st stage - Tier 1 experiments

During the Tier 1 experiments, we conducted a post-mortem analysis on reproducibility of ex situ samples in WP5 laboratories. This was introduced in the D5.5 deliverable (Section 2.5.2), and further discussed in the D5.3 deliverable (Section 2.1.2). To summarize: (1) Research partners prepare and cycle coin cells following BIG-MAP WP8 protocols. (2) Cells are sent to experimental partners. (3)

[13] Linear fitting is sample-independent, Shirley is based on the estimation of the inelastic scattering of electrons as they move through the solid state using the Shirley model, and Tougaard incorporates the concept of inelastic energy loss cross-sections to describe the background signal.

Receiving partners perform uncoordinated measurements. (4) They also perform independent data analysis, using their preferred algorithms, software and protocols.

We established a specific framework for XPS data acquisition and analysis, as illustrated in Figure 12. For clarity, we chose to only focus on samples prepared at CTH. In this scheme, samples (aged cells with lithium and graphite as electrodes – Li//Gr – and graphite and LNO as electrodes – LNO//Gr) were sent to CEA, UU and SOLEIL. Every institution conducted measurements of the electrochemical data for the sample cells. There was no direct communication between receiving institutions or adherence to a specific protocol during the experimental phase, resulting in uncoordinated execution of the XPS experiments over a period of months. Additionally, the UU data were obtained using charge neutralization, which was not the case for CEA and SOLEIL data. Subsequently, data analysis was independently conducted at each institution by individuals possessing varying levels of expertise in XPS and/or batteries.

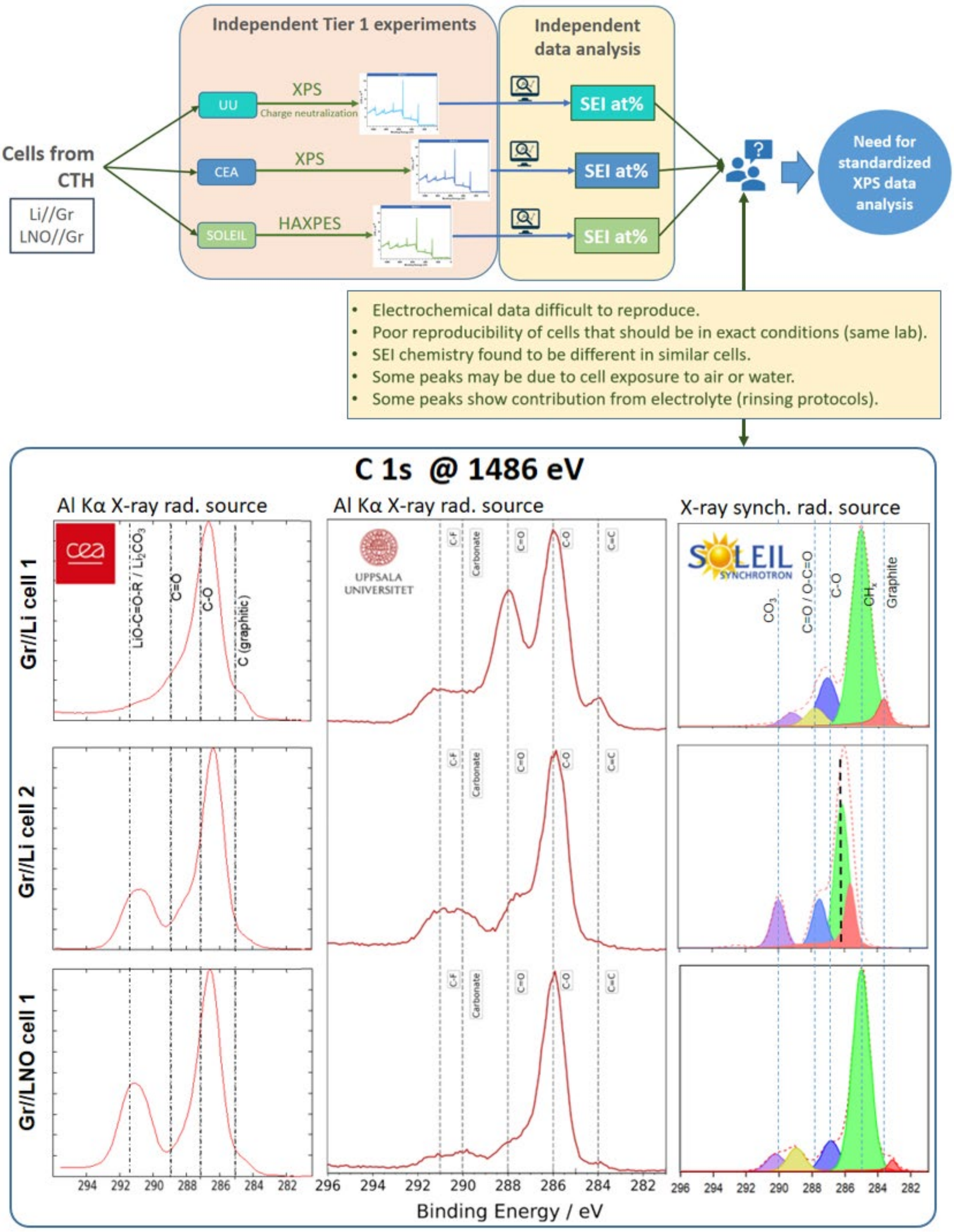
We encountered challenges in reproducing the electrochemical data consistently, as well as poor reproducibility of the experimental results for cells that should be in exact conditions since they were assembled in the same laboratory. Moreover, when comparing the data analysis performed by the different partners, discrepancies arose in the identification of the SEI composition. We also identified some peaks resulting from cell exposure to air and/or water and other peaks from electrolyte contribution.

Discrepancies can be attributed to multiple factors, including:

- Variations in cell assembly and cycling conditions.
- Differences in the timing of the XPS experiments. They were conducted at different time intervals after receiving the cells.
- All samples were handled with basic precautionary measures, such as the use of a glove box for electrode storage and a transfer vessel to prevent air exposure.
- The use of diverse data acquisition protocols (pass energy, use or not of charge neutralization).
- The choice of data analysis software plays a minor role since software packages are based on the same algorithms for peak fitting.
- Human factors can significantly influence the results. For instance, the expertise of the person analyzing the spectra plays a role in the observables.

This initial analysis highlights the importance of establishing standardized protocols for both data acquisition and analysis in XPS experiments within the context of BIG-MAP. Reducing errors caused by these various contributions is crucial. Within our WP, we have identified three key aspects that require attention to standardize XPS data analysis:

- **Cell contribution**: Consistency in cell preparation and cycling is key. Electrodes should undergo uniform rinsing by a designated partner, and thorough documentation of cell production and aging processes is essential.
- **Experimental contribution:** To enable comparisons between different partners, XPS measurements should be performed on cells with the same age which underwent the same cycling history.
- **Data analysis contribution:** Adherence to specific protocols for data analysis is needed to ensure consistency and accuracy.



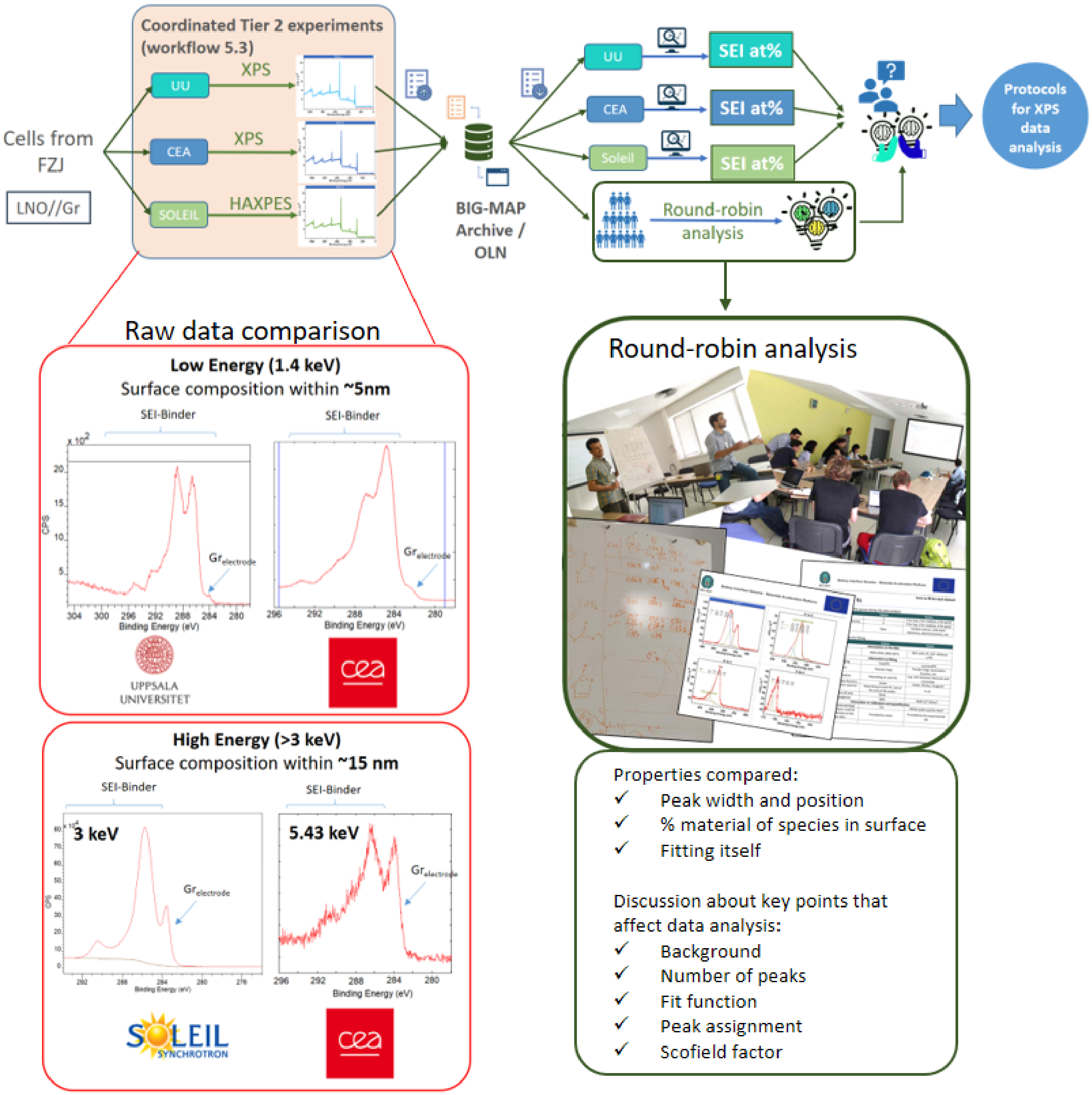
- Electrochemical data difficult to reproduce.
- Poor reproducibility of cells that should be in exact conditions (same lab).
- SEI chemistry found to be different in similar cells.
- Some peaks may be due to cell exposure to air or water.
- Some peaks show contribution from electrolyte (rinsing protocols).

Figure 12. Organization and results for the post-mortem XPS data analysis performed during Tier 1 experiments. Batches of Li//Gr and LNO//Gr cells prepared at CTH were sent to CEA, UU and SOLEIL, where independent experiments and data analysis were done. At the bottom, C 1s core level peaks and analysis comparison on three cells. The comparison of the results yielded inconclusive findings, standardizing experimental conditions and data analysis is necessary.

Building upon the experience and results obtained from the Tier 1 experiments, a second approach was developed. The objective was to evaluate the cell manufacturing and experimental processes, and to find protocols for standardized XPS data analysis. This process was conducted within the framework of the 5.3 experimental workflow during Tier 2 experiments.

3.3.2 Post-mortem analysis 2nd stage - Tier 2 experiments under the 5.3 experimental workflow

Tier 2 experiments were defined during the first semester of 2022 and launched in September 2022. All experiments in the 5.3 workflow were designed to use coin cells prepared by a unique provider using the same protocols. Figure 13 displays the specific workflow for the XPS data analysis (see D5.3 deliverable, Section 3 for more details in the experimental procedure).



Properties compared:
- ✓ Peak width and position
- ✓ % material of species in surface
- ✓ Fitting itself

Discussion about key points that affect data analysis:
- ✓ Background
- ✓ Number of peaks
- ✓ Fit function
- ✓ Peak assignment
- ✓ Scofield factor

Figure 13. Scheme of the post-mortem XPS analysis carried out during Tier 2 experiments under the 5.3 experimental workflow. Comparison of the raw data on graphite obtained at UU, CEA and SOLEIL show comparable findings. A round-robin analysis was done to perform and discuss data analysis.

Details of the XPS workflow are the following: (1) Batches of LNO//Gr coin cells, with and without LiTDI as electrolyte additive, were prepared by FZJ (Dec 2022), following unique protocols for assembling, cycling, washing and handling. (2) Different samples were sent at the same time to the three partners: UU, SOLEIL and CEA. (3) Partners performed coordinated XPS measurements, within a few weeks of receiving the samples. (4) Data is uploaded to the BIG-MAP archive. Information of the experiments and metadata uploaded to the BIG-MAP Notebook when possible. (5) Expert XPS partners performed independent data analysis using their preferred tools, algorithms, software and protocols. (6) An in-person XPS data analysis workshop was held, during which a round-robin data analysis exercise was carried out to compare results and comment on XPS data analysis protocols. (7) Experts collaborated to analyze data analysis results and to establish protocols for XPS data analysis.

3.3.3 Round-robin data analysis

In June 2023, WP5 and other invited BIG-MAP colleagues met in Grenoble, France, to participate in an XPS data analysis workshop with the objective of training BIG-MAP scientists on XPS data analysis, while raising awareness about the need for standardization in data analysis. This was centered around a round-robin XPS data analysis, bringing together expert and non-experts on XPS, using the same software tools to analyze the same data simultaneously. Discussions focused on caveats, challenges and the needs to achieve standardization in XPS data analysis. A critical aspect in XPS data treatment is the quantification of elemental concentrations. The peak fitting model and the sensitivity factor has to be carefully evaluated for more accurate semi-quantitative analyses. Therefore, the second aspect of the round-robin analysis was to test the reproducibility of quantitative results from data treatment performed by experts from different laboratories.
At the end of this section, building on these aforementioned exercises, we will provide XPS users with recommendations for XPS data standardization.

- **Participants**

Thirteen people participated in the XPS workshop, including three XPS experts who tutored the hands-on session and round-robin analysis. The list of participants can be found in Annex 9.3.

- **Data**

We used data from the 5.3 experimental workflow: XPS spectra recorded on a graphite electrode by CEA using Al Kα X-ray source, measured in both survey and high-resolution modes. Two samples were used, BM5-034 and BM5-037. These correspond to aged coin cells with and without LiTDI as electrolyte additive, respectively. For this practical session, we focused on the graphite electrodes.
Data reference: https://archive.big-map.eu/records/cs4t5-e8823

- **Report form**

During the hands-on session of the workshop, participants were requested to complete a report form. They provided information on their background in XPS and battery chemistry and indicated their field of expertise. The report also collected information on the background and peak fitting, including selected background and fitting functions, statistics, and information on calibration and quantification. Participants recorded the results of the fitting for each core level: position in eV, FWHM in eV and peak intensity in counts/s. An example of the form is presented in Annex 9.3.

- **Minutes of the workshop**

1. Introduction to the XPS technique and data analysis.
2. Participants downloaded datasets from the BIG-MAP archive.
3. Inspection of the survey mode spectra: this revealed that the two samples share some peaks, while others are not shared.
4. Tutors demonstrated how to use CasaXPS and explained the basic method for data analysis, including energy calibration, scale factor adjustment, background fitting and peak fitting.

5. Hands-on session, where participants worked alone or in pairs to analyze the high-resolution mode spectra and fill out the report form.
6. Several handbooks and databases including a selection of review papers were available as support for all participants.
7. Round-robin analysis, each participant or team presented their results, arguing their choices.
8. Participants sent their fittings and reports to tutors for further analysis.

Methodology

This unique round-robin data analysis emphasizes the critical aspects of accurate peak fitting, from both mathematical and chemical perspectives.

Herein, we present a comparative analysis of blind XPS data treatment results obtained by three participants having different backgrounds, labeled as Participant-A, -B and -C. Participant-A used linear background subtraction function while Participants-B and C used Shirley background (most used one in the literature). There were also differences in the chosen fitting function. Table 2 lists the main information for these participants.

Table 2. Background of participants and general fitting information on exemplified participants to the XPS round robin analysis.

	Participant-A	Participant-B	Participant-C	Notes
Knowledge on XPS	1	1	2	*1 for low, 2 for medium, 3 for good*
Knowledge on battery chemistry	2	3	2	*1 for low, 2 for medium, 3 for good*
Field of expertise	Data	Solid state chemistry	Surface science	*Surface science, solid state chemistry, electrochemistry*
Background function	Linear	Shirley	Shirley	*Linear, Shirley, Tougaard*
Fitting function	Pseudo-Voigt	Lorenztian Asymmetrical	Pseudo-Voigt	*Pseudo-Voigt, asymmetric function, etc.*

All participants had two sets of XPS core level spectra of F 1s, C 1s, N 1s, O 1s, P 2p and Li 1s for graphite cycled with and without LiTDI based electrolyte used in BIG-MAP (LP57). To obtain accurate peak treatment, we use the F 1s and compare the results from the three participants from Table 2. Figure 14 displays the peaks fitted by the three participants. the three participants.

Peak fitting comparison

In the case of the F 1s spectrum recorded at the surface of graphite cycled without LiTDI, two distinct peaks are identified centered at 685 and 687 eV (see Figure 14) respectively assigned to LiF and PF_6 ($LiPF_YO_Z$) groups. Considering only a mathematical perspective, it is indeed possible to use two peaks to fit the experimental data. All participants did so and accurately performed peak fitting. The only differences in the analysis methodology stem from the choice of fitting functions. For background treatment, some participants opted for a linear function instead of a Shirley function. For peak fitting, some participants chose asymmetric functions instead of the pseudo-Voigt one. While these choices do not significantly change the final interpretation in our example, they may impact the fitting in the case of metal-related core-level spectra where satellite structures are usually present near the main photoemission structure (herein indicating that the support of an XPS expert is mandatory for data analysis).

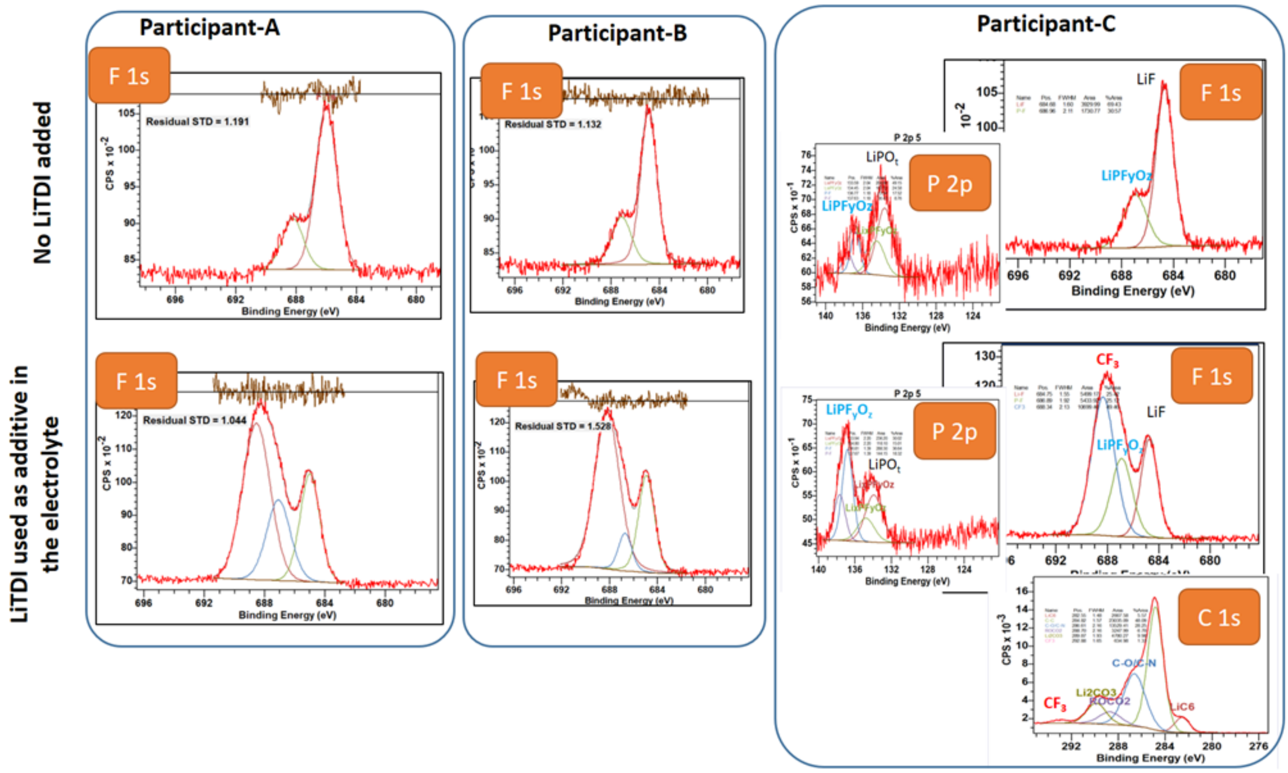


Figure 14. Comparison of the fitted F 1s core level for three chosen round-robin analysis participants.

In the case of the cell with LiTDI (molecule containing CF_3 group), the peak at 687 eV shows an asymmetry toward lower binding energy with a peak broadness, signature of both a CF_3 related group in LiTDI and a PF_6 group. In this case, differences in peak fitting were observed among all participants. This is a clear example highlighting how mathematical considerations are not sufficient to describe the experimental data. Only the battery expert group, Participant-C, used stoichiometric considerations using the information from P 2p and C 1s core level spectra. Indeed, the CF_3 signature can be detected in C 1s (peak at 293 eV) and PF_6 in P 2p (peak at 136 eV) spectra. Therefore, the best fit in this case is to consider simultaneously F 1s, P 2p and C 1s core level peak by taking into account peak position, stoichiometry ratio and FWHM values (Figure 14). In this case, the best XPS fit is always intimately related to the electrochemical process. In this particular case, the F 1s peak is the signature of salt or additive degradation by-products according to the following the electrochemical reactions:

$$LiPF_6 \rightarrow PF_5 + LiF$$
$$LiPF_6 + Li_2CO_3 \rightarrow 3LiF + POF_3 + CO_2$$
$$PF_5 + Li_2CO_3 \rightarrow 2LiF + POF_3 + CO_2$$

The presence of F 1s peaks at 685 and 687 eV, and P 2p peaks at 133 and 136 eV, are the signatures of the by-products resulting from the reaction path above. Herein demonstrating that further physic-based knowledge is needed to fine-tune the purely mathematical spectra peak fitting.

Conclusions and final recommendation for XPS data standardization

In this original round-robin study, we highlighted the importance of respecting specific protocols at each stage of the analysis to achieve the most accurate peak fitting. These stages include peak calibration, background subtraction, peak profile distribution, number of peaks, and FWHM, peak position and peak intensity evaluation. The most critical challenges occur when dealing with asymmetric peaks with a FWHM larger than 3 eV. In such cases, cross-linking information from different peaks proves to be the most effective approach for building a fit that provides insights into the true surface chemical distribution.

It is evident that building an XPS peak fitting database would significantly enhance the accuracy of peak fitting, particularly in the context of complex battery chemistry. This would therefore form an important objective for future projects.

The results of the round-robin analysis will be presented in a forthcoming publication, providing additional information in terms of quantification and further depth and insights into the implications of our research.

4 Use case: Vibrational spectroscopies - Raman and Fourier transform infrared

4.1 The scientific motivation

Raman and Fourier Transform Infrared (FTIR) are commonly employed, complementary vibrational spectroscopy techniques. They are used to probe vibrational excitations of the sample and provide unique structural fingerprints for molecule identification (see Annex 9.1 for more details). Raman is widely used for bulk and near-surface operando characterization of battery materials - covering anode/electrolyte and cathode/electrolyte interactions. FTIR is mostly used to probe the electrolyte solvation structure and speciation, as well as electrode/electrolyte interfaces.

The surface sensitivity of both Raman and FTIR can be enhanced with specific strategies [17]. During Tier 1 experiments, we explored them for the attenuated total reflection IR, performed at CTH, and for the diffuse reflectance FT spectroscopy, performed at ICMAB, to identify the SEI on cycled graphite electrodes. However, it was recognized during these Tier 1 experiments that the sensitivity of the techniques was insufficient to probe the SEI formed on graphite electrodes after the formation cycles. Thus, for Tier 2 experiments, we focused on vibrational spectroscopies for the analysis of LNO//Gr cell bulk properties.

4.2 Raman and FTIR in experimental workflow

Figure 15 displays the vibrational spectroscopies from LP57 electrolyte (1M $LiPF_6$ in EC:EMC 3.7 wt%), as well as LNO and graphite electrodes. Panel a) in the figure shows the Raman and FTIR spectra of LP57, where characteristic vibrations, such as C-H and C=O stretching modes, EC ring breathing and P-F stretching of PF_6^- anion are highlighted. Panel b) gives the Raman spectrum of LNO and its two characteristic vibrational modes are identified: the stretching (E_g) and bending (A_{1g}) modes. Panel c) shows the spectrum of the BIG-MAP composite graphite electrode (graphite active material mixed with high surface disordered carbon conductive additive) provided by CID. This spectrum profile consists of 5 peaks: G, D1(D), D2, D3 and D4, where the G band is characteristic of the in-plane bond-stretching of pairs from C *sp*^{2} atoms. D1 and D2 are attributed to the vibration mode of disordered graphitic lattice (A_{1g} and E_{2g} symmetry). D3 and D4 are linked to the presence of amorphous carbon, while D4 is also related to the presence of C-H termination group [18-22].

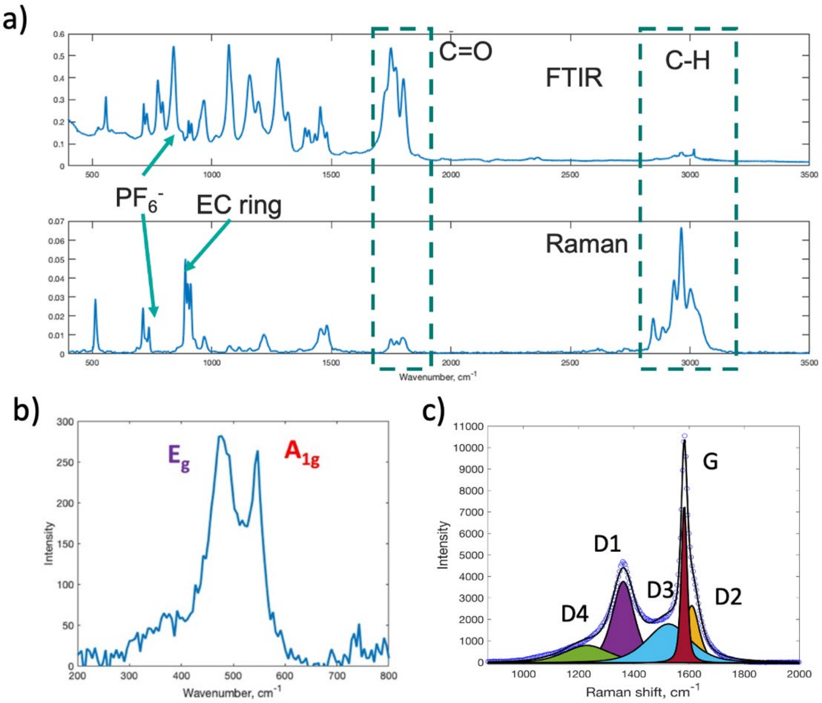


Figure 15. Vibrational spectroscopies of BIG-MAP materials: a) FTIR and Raman spectra of LP57 electrolyte; b) Raman spectrum of LNO; c) Raman spectrum, with its deconvolution, of the graphite composite electrode from CID. Spectra were measured at CTH.

4.3 Implementation of data analysis standardization

Similar to other spectroscopic techniques, the analysis of Raman and FTIR spectra consists of three steps: trimming, baseline subtraction and peak fitting (shown schematically in Figure 16). The implementation of each of these steps can give a slightly different output depending on the employed processing software and user expertise. The accumulation of small deviations during each processing step may result in different final outputs. In order to achieve the most reliable and reproducible data analysis it is therefore essential to define standards and protocols for individual processing steps, including employing the same processing. For this purpose, a collaboration between WP5 and WP9 has been crucial and has resulted in the development of the PRISMA software, which is particularly useful for both standardization and automation of vibrational spectral data analysis.

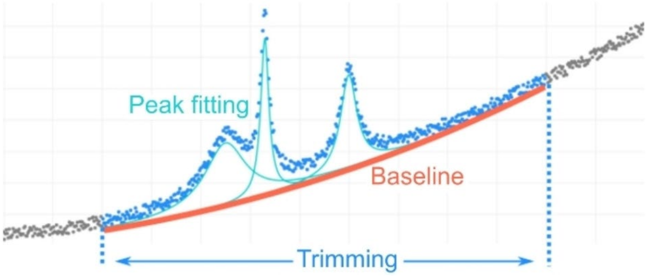


Figure 16. The three steps of spectral analysis: trimming, baseline fitting and subtraction, and peak fitting. Extracted from Flores et al. [4].

Step 1. Data trimming. Experimental Raman and FTIR data are typically recorded within a wide wavenumber range, typically spanning from 0 to 4000 cm^{-1}. While keeping a complete wavenumber dataset is important for future data analysis, the spectra should be limited to the specific wavenumber range of interest before performing background subtraction and peak fitting. The best results are achieved through data trimming, ensuring that only relevant peaks are included, along with additional wavenumber intervals (both at lower and higher values) relative to the peaks of interest to facilitate accurate background estimation. Including wider wavenumber interval may lead to a less accurate background estimation due to the non-uniformity of the background across the entire spectrum. Conversely, trimming too close to the peaks of interest may incorporate part of the peak tales into the background contribution, thus altering the peak profile. Optimal data trimming requires visual assessment, however, a range of +/- 100 to 200 cm^{-1} framing both sides of the peaks of interest is typically sufficient for accurate data analysis and interpretation. Examples of appropriate data trimming are shown in Figure 15 b) and c).

Step 2. Background subtraction. Background removal is a very important step in vibrational data analysis; that can only be correctly achieved on appropriately trimmed data (step 1) and is essential for accurate peak fitting (step 3). While there are numerous methods and software available for background subtraction, the AsLS methods was chosen within the BIG-MAP project (see Section 2.2.2) due to its simplicity of use and rapid computation performance (the last being a key factor for high-throughput analysis). We follow the same workflow as for XRD data (Section 2.2.2), by tuning the p and λ parameters.

Step 3. Peak fitting. Finally, the peak fitting provides the most valuable information pertaining to material structure and composition. The number of fitted peaks must be equal to the number of Raman or IR active bands in the spectrum. The extracted information from peak fitting is: peak positions, peak intensities and FWHM (later referred as width). In turn, peak positions give information regarding the chemical bonds in the sample, intensities provide material quantification and insights into polarizability/dipole moment and, lastly, peak profile and width are related to the sample crystallinity/disorder. The peak profiles included in PRISMA describe the most common bands, which are either naturally (Lorentzian), partially (pseudo-Voight) or heavily (Gaussian) broadened by the measurement conditions [23, 24].

To perform the peak fitting using PRISMA, the user must indicate the number of peaks, the fitting function, the lower and higher bound for each of the peak positions and the maximum peak width. The code then fits the complete spectrum as a sum of peak profiles using the Trust Region Reflective Algorithm as implemented in the SciPy package [25]. By applying the above-mentioned process, PRISMA-assisted peak fitting ensures a high degree of data analysis standardization.

4.4 Implementation of data analysis automation

As demonstrated, PRISMA is a powerful tool for analysing vibrational spectroscopy data. In addition to individual spectral analysis, PRISMA offers automated data analysis for a series of dependent spectra, i.e., spectra with the same set parameters for trimming, background subtraction and peak fitting, through a customized Python script that is used through its GUI. The spectra as a function of a perturbation variable (e.g., time, temperature, coordinate, etc.) are provided as an input variable to the script. Trimming, background and peak profile parameters are manually optimized for a representative number of spectra and then fixed for all the datasets. Specifically, background subtraction is performed using an AsLS function with fixed parameters p and λ, while the number of peaks is defined by the user from the expected composition and chemical nature of the sample. Each spectrum is automatically fitted with the same selected parameters. Chi-square was used as a metric to assess the fit quality. Processed spectra and peak parameters (position, width and intensity) as a function of the perturbation variable are recorded in an output file.

The typical workflow when using PRISMA to analyse a sequence of dependent patterns is the following:

1. Selecting a pipeline, that is, a pre-defined recipe of spectrum processing steps.
2. Upload of the input pattern file(s). PRISMA offers three parsers able to load patterns in specific formats: several individual .txt files, a single .txt file, and a single .csv file.
3. Selection of a spectrum for visualization.
4. Tuning the processing parameters until reaching a visually satisfactory fit. For baseline correction, the parameters are the trimming interval and the λ and p parameters of the fit; for peak fitting these are the number of peak profiles, their bounding neighbourhood, and a maximum width limit.
5. Inspection to verify whether the parameters enable satisfactory fits of other patterns or if they need further tuning.
6. Iteration between steps 4 and 5 by selecting multiple patterns and visually inspecting the results, until finding satisfactory parameters that apply to most patterns.
7. Running a high-throughput processing of all patterns with the chosen parameters.
8. Downloading the results for further analysis and plotting.

As part of the development and evaluation of the PRISMA software, two Raman use cases were chosen: Raman spectroscopy on ethylene carbonate (EC) monitored during its melting transition and Raman mapping of a composite graphite electrode. Here we summarize the main results, which are published in Flores et al. [4]. In the first case, the temperature-dependent spectra were processed with PRISMA, where baseline correction and peak fitting were performed. The resulting peak positions and intensities as a function of temperature provided valuable insights into the structural rearrangement of EC upon melting. For the second example, Raman mapping analysis of graphite and carbon additive distributions in a composite graphite electrode, PRISMA multi-peak fitting model assessed chemical heterogeneities, graphitic regions, and the presence of both disordered and amorphous carbon. The subsequent analysis demonstrated the value of studying peak trends instead of raw spectra, and PRISMA's interface facilitated analysis parameter tuning and results visualization.

Operando Raman of LNO cathode. After the success of these two use cases mentioned above, PRISMA was used to perform automated operando Raman peak analysis of LNO, focusing on fitting the Ni-O stretching (A_{1g}) and Ni-O-Ni bending (E_g) bands (see Figure 17) during charge and discharge cycles. The fitting process provided valuable peak parameters, including peak positions, widths and intensities. The data analysis provides insights on the evolution of the E_g and A_{1g} bands during cycling, contributing to the understanding of the behavior and transitions of LNO and other Ni-rich materials. This work will be soon published as a result of an inter-WP collaboration.

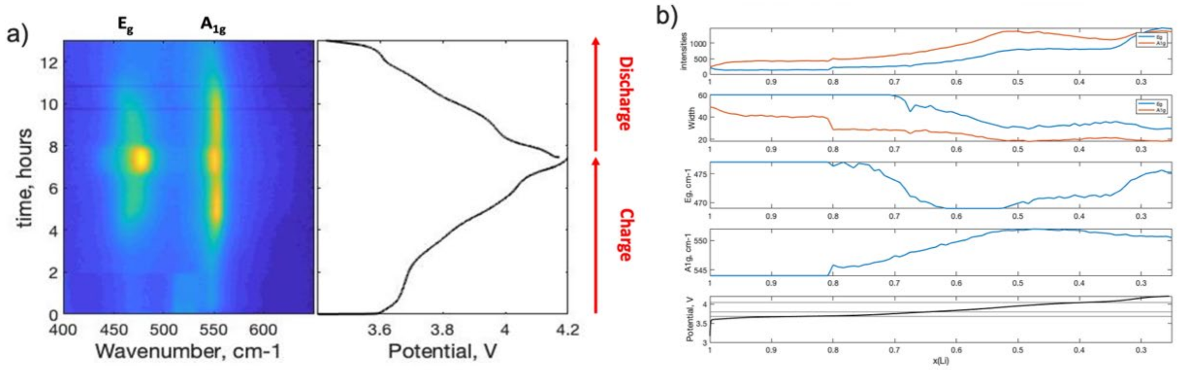


Figure 17. a) Operando Raman spectra of LNO during first charge-discharge cycle. b) output results of two peak PRISMA assisted fittings: A_{1g} and E_g bands positions, intensities, and widths evolution as a function of the lithiation degree.

4.5 Implementation of data analysis correlation with modelling

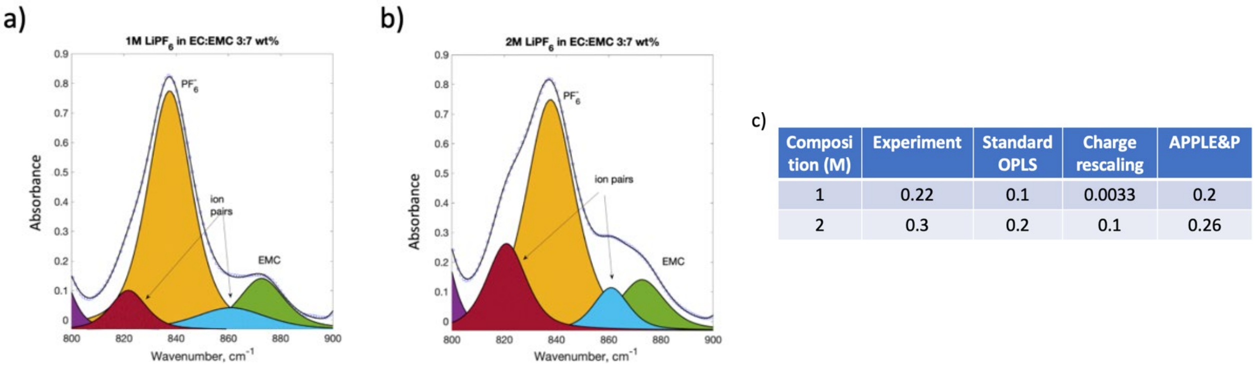
Data analysis correlations between experimental and computational data are important to validate the theoretical models. This implies the selection of common measurable/computable sample properties. Then, theoretical models impacting this property can be developed, and experiments can be carried out to establish this correlation.

A successful BIG-MAP collaboration has been established between partners from CTH, WWU and FZJ to cross-correlate molecular dynamics simulations performed in WP3 with experimental results from WP5. This work has been recently published in Maiti et al. [26].

Sample: Two Li-ion battery electrolytes: 1M (and 2M) $LiPF_6$ in EC:EMC 3:7 wt% (LP57)
Sample property to correlate: Coordination numbers (CNs)
Theoretical models: Molecular dynamics (MD) simulations
Experimental data: FTIR

MD simulations were performed using three force fields: polarizable force field (APPLE&P model), standard non-polarizable force field without charge rescaling (standard OPLS model), and standard non-polarizable force field with charge rescaling (charge rescaling model). They were compared for their efficiency to model the structural properties of two Li-ion battery electrolytes. Specifically, electrolyte solvents and anion coordination statistics with respect to the electrolyte cation were calculated and compared to the experimental values obtained from FTIR spectra.

Experimental FTIR data was processed with PRISMA to extract the main observables. Figure 18 a) and b) displays the data and peak fitting. The FTIR spectra have a characteristic P-F bond stretching mode at 838 cm^{-1} and an EMC peak at around 873 cm^{-1}. The increase in $LiPF_6$ concentration results in an increase of the two shoulders centered around 820 and 860 cm^{-1} (both blue- and red-shifted as compared to free PF_6^- peak), that have been attributed to the presence of Li^+–PF_6^- ion pairs/aggregates [27-30]. The specific CNs of Li-ions around fluorine of PF_6^- ions were calculated for two salt concentrations, according to the formula $CN = A_C/(A_F + A_C)$, where A_F is the peak area of free PF_6^-, and A_c the peak area of PF_6^- coordinated to Li^+ ion pairs. The CN values extracted from experiments and MD simulations are shown in the table in Figure 18 c). We find that there is a very good agreement with the polarizable force field model. By contrast, the OPLS force field underestimates the cation–anion pair formation. It is notable that, after charge rescaling of the OPLS force field model, the structural agreement becomes even poorer.



Composition (M)	Experiment	Standard OPLS	Charge rescaling	APPLE&P
1	0.22	0.1	0.0033	0.2
2	0.3	0.2	0.1	0.26

Figure 18. Correlation between FTIR experimental spectra and MD simulation. FTIR spectra of P–F stretching regions of a) 1 M LiPF$_6$ in EC : EMC 3 : 7 wt% (LP57) and b) 2 M LiPF$_6$ in EC : EMC 3 : 7 wt% electrolyte and corresponding peak fitting. c) Comparison of coordination numbers of Li$^+$ cation with PF$_6^-$ anion. The compositions are slightly different in simulation (EC : EMC = 24 : 76) from experiment (EC : EMC = 30 : 70). Adapted from Maiti et al. [26].

This example showcases the importance of validating theoretical models with high fidelity experimental results processed through appropriate analytical software.

5 Use case: Tomographic segmentation through a machine learning approach

Here we outline an automated approach to big data analysis using, as an example, results obtained by ex situ and operando X-ray tomography.

5.1 Motivation and segmentation for imaging data

In general, battery electrodes have a complex 3D microstructure, with a near-random spatial organization, which in turn effects their operational properties. Moreover, during cycling, dynamic changes occur in the material's microstructure. These changes can be investigated with non-destructive 3D imaging techniques, such as X-ray tomography, which produce large experimental datasets at each acquisition/time-step to obtain final 3D volumetric reconstructions with sufficiently high resolution.

Standard 2D images are composed of an array of unit elements, the pixels. Similarly, 3D volumes are composed of regular cubic elementary volumes called voxels. In a raw 3D volume obtained through tomography, each voxel has a value which can be linked to a material in the sample. Segmentation is the process of attributing a phase to each voxel in the raw volume. This can be a rather complicated process depending on the imaged materials, the acquisition resolution, image quality and the fidelity expected in the analysis. Quantitative analysis requires the segmentation to be precise as this strongly influences the ultimate analysis precision and fidelity. Thus, segmentation of tomographic datasets for quantitative analysis is then a long and challenging process.

For complex microstructures, standard segmentation algorithms tend to fall short when aiming for the highest fidelity; machine learning models can then be investigated to improve this. Best results seem to come from either highly specific algorithms or U-Net [31] like based CNNs, both of which are very time consuming, human intensive and require specific setups. When considering the training of CNNs for segmentation, image annotation is the most time consuming and human intensive step as it has to be done manually. In order to minimize the bottleneck nature of this step, we developed, in collaboration with WP11 and DTU (L. Rieger), a segmentation tool making use of *active learning* during the training section in order to optimize the human time allocated to manual annotation.

5.2 Method description

Usually, U-Net models for segmentation require a certain amount of annotated data for a complete training. For one 3D volume from X-ray tomography (with dimensions of 2048x2048x2048 voxels), between 5 to 10 full slices of the volume (i.e., 5 to 10 2048x2048 images) must be manually annotated. Depending on the user and complexity of the material/sample, this can take several days. Designing our segmentation tool, we chose a lighter strategy for the initial annotated training input by partially annotating full slices from the volume (also between 5 to 10). This process takes between 1 to 3 hours. Once this step is completed, the model hyper parameters are set and the training is started.

During training, the model checks the metrics (validation score) after a fixed (parametrized) number of epochs. If the validation score does not improve on the last few epochs, the training process stops. Once this initial training step is completed, the network examines an unannotated pool of patches from slices (typically the whole volume or a subset thereof) and identifies the most suitable patches to add to the training dataset. Then, a predefined number of patches is outputted by the network, including both predictions and input images. These patches are meant to be manually reannotated for the training to continue. A portion of this new training data is automatically used for training and another for validation. These patches are also automatically removed from the unannotated pool. This annotation step needs to be thorough, performed on the entire patches which are usually encompassing complex areas. Patches are significantly smaller than the full slice, thus the annotation process is relatively fast, up to 1 h on 5 patches. Additionally, speeding up the manual annotations are the network predictions, they serve as a starting point that only requires correction rather than starting from scratch. Once the new patches have been reannotated, the training can resume, following the same previously mentioned workflow, i.e., requesting user annotation for a new set of patches when necessary. This process can be repeated until the user is satisfied with the outputs provided by the network or until the global total number of epochs allowed for training is reached. Either way, the full volume segmentation can then be performed. Figure 19 illustrates the schematic view of this segmentation tool.

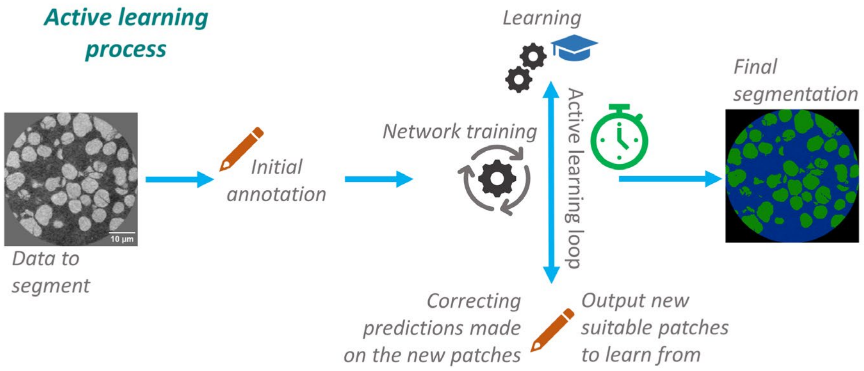


Figure 19. Schematic view of the active learning accelerated training process proposed with the new segmentation tool developed in BIG-MAP. Orange pen markers are identifying the two manual annotation parts of the workflow. The blue graduation cap indicates where the learning is made for the network in the active learning loop.

5.3 Practicalities and tool availability

This tool was developed using Python and the PyTorch module through scripts without GUI. For the work performed in BIG-MAP, the Krita[14] software was used to make the manual annotations. Krita is a free, open source software designed for digital painting, that can be scripted using Python. Additional scripts and plugins were developed to help the import/export process between the network environment and the Krita working environment. Any alternative software can be used as long as it has (1) the capabilities to load the network outputs, (2) annotation can be performed using a pen tablet and (3) that its outputs are compatible with the network input format.

At the moment of redaction of this report, this is work in progress. While a functional prototype of the entire process has been achieved, we are now focusing on the tool optimization and stability. A publication is in preparation and the source code for the active learning segmentation tool will be published on GitHub. The different scripts and plugins used to include Krita in the manual segmentation process will also be made available on GitHub at the same time. The code publication will also be accompanied by a tutorial.

6 Guidelines for analysis of characterization data

Our data analysis work on the multi-structure, multi-scale characterization data allows us to identify guidelines for data analysis standardization and automation.

6.1 Data analysis standardization guidelines

Data analysis standardization of characterization techniques can be achieved by following defined protocols and workflows. Figure 20 displays a step-by-step guide to follow to formulate and implement protocols for data analysis standardization.

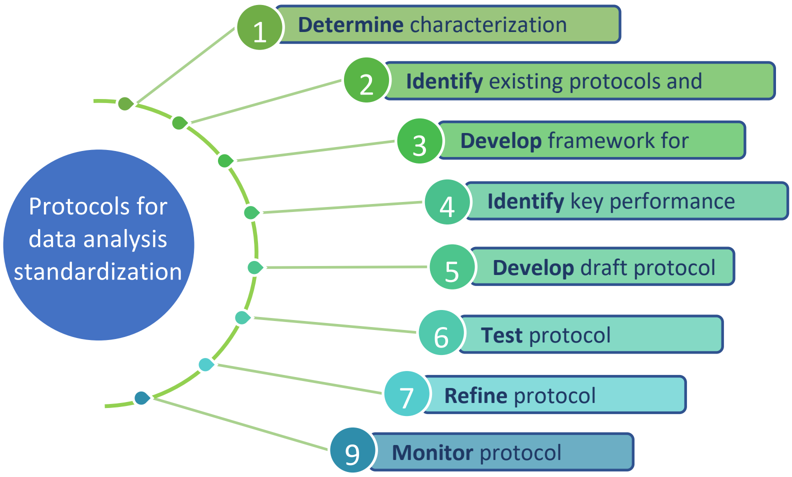

Protocols for data analysis standardization
1 Determine characterization
2 Identify existing protocols and
3 Develop framework for
4 Identify key performance
5 Develop draft protocol
6 Test protocol
7 Refine protocol
9 Monitor protocol

Figure 20. Scheme illustrating protocols for data analysis standardization.

After selecting the characterization technique for standardization, it is essential to identify existing standards and protocols while analyzing gaps. Based on this, we can develop a protocol framework outlining the steps to be followed for data analysis and reporting. Key performance indicators for the protocol, such as accuracy, precision, and reproducibility, need to be identified. Subsequently,

[14] Krita: https://krita.org/en/

a draft protocol with detailed data analysis instructions can be created, tested on a small dataset, refined based on feedback, and finalized. Regular monitoring and protocol updates are also necessary to adapt to evolving characterization techniques.

6.2 Data analysis automation guidelines

Figure 21 displays our proposed step-by-step guide to automate data analysis.

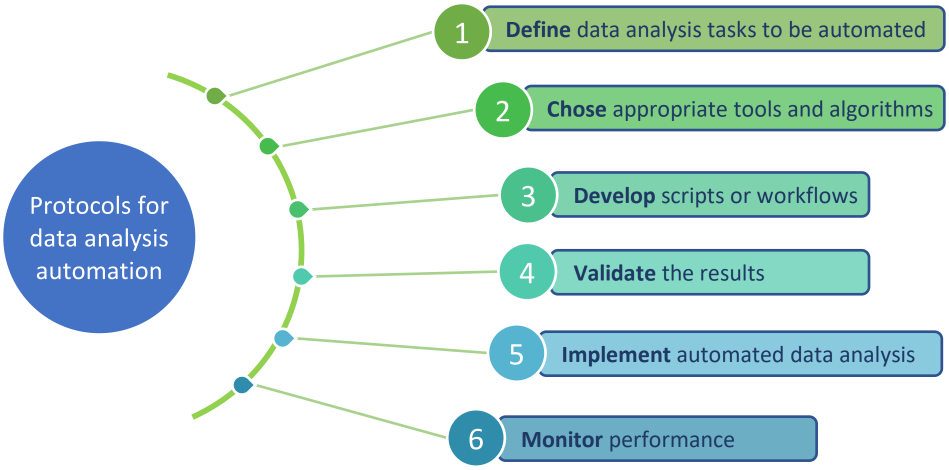
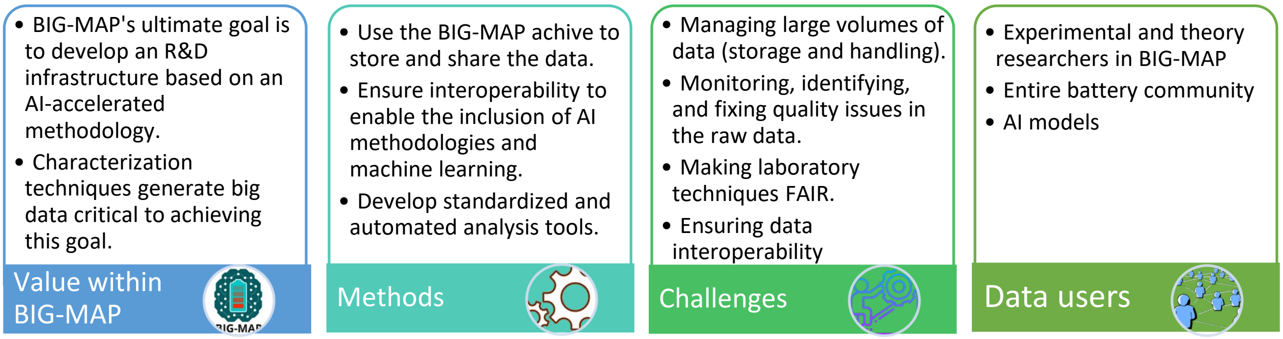

Protocols for data analysis automation
1 Define data analysis tasks to be automated
2 Chose appropriate tools and algorithms
3 Develop scripts or workflows
4 Validate the results
5 Implement automated data analysis
6 Monitor performance

Figure 21. Scheme illustrating protocols for data analysis automation.

We start by identifying data analysis tasks for automation (e.g., data pre-processing, data fitting, feature extraction and formatting). We then select suitable tools and algorithms for each task, e.g., Python has many specific libraries to handle datasets and perform analysis. Scripts or workflows to automate analysis tasks will be developed and validated against manual analysis or existing benchmarks to ensure accuracy and reliability. Next steps are to implement the proposed automation protocol and monitor its performance over time, enabling the user to produce updates with evolving data and techniques.

7 Big data handling and data FAIRness

Advancements in battery characterization technologies have generated increasingly diverse and complex datasets, often with significantly large volumes. This situation characterizes what is commonly referred to as "big data". Traditional processing techniques are often insufficient, as they result in slower processing and analysis. Hence the need for flexible software tools for analysis, deploying novel methods and algorithms to extract valuable insights [32].
In our team, we labeled big data as voluminous data sets that cannot be handled individually, with diverse data types (structured, semi-structured, and unstructured), often generated at high speed.



Value within BIG-MAP

- BIG-MAP's ultimate goal is to develop an R&D infrastructure based on an AI-accelerated methodology.
- Characterization techniques generate big data critical to achieving this goal.

Methods

- Use the BIG-MAP achive to store and share the data.
- Ensure interoperability to enable the inclusion of AI methodologies and machine learning.
- Develop standardized and automated analysis tools.

Challenges

- Managing large volumes of data (storage and handling).
- Monitoring, identifying, and fixing quality issues in the raw data.
- Making laboratory techniques FAIR.
- Ensuring data interoperability

Data users

- Experimental and theory researchers in BIG-MAP
- Entire battery community
- AI models

Big data in science and in WP5

The significance of big data in science originates from the continuous evolution of science, marked by significant advancements in scientific instruments and simulations that generate large volumes of data. Big data allows to conduct comprehensive data exploration, investigating large and diverse datasets to unveil patterns and correlations that were previously inaccessible. Fundamentally, this will improve and accelerate R&D by the development of automation of data processing, analysis and visualization, streamlining workflows and delivering faster, innovative results. Big data also promotes the principles of data FAIRness (Findable, Accessible, Interoperable, and Reusable) [33], fostering reproducibility and transparency in scientific research.

In WP5, we have identified various techniques that generate big data. These include neutron tomography (conducted at ILL by CEA), X-ray tomography (performed at ESRF by CEA and CTH), and operando experiments on NPD (at ILL) and XRD (at DTU and ESRF). Imaging and tomography techniques produce large volume datasets, while operando experiments for battery cycling generate hundreds or thousands of files. Some of this data originates from LSFs, where appropriate servers exist to store both raw and processed data. Other data is generated at laboratories, typically comprising a large number of spectra files that are produced at high speed during operando characterization experiments. These files can be stored on local laboratory PCs and uploaded to the BIG-MAP Archive.

Data analysis workshop: learning best practices for data analysis and big data

On June 16th, 2023, we conducted a workshop on big data handling and analysis in WP5. This workshop helped scientists to identify whether they are dealing with big data, enhancing their awareness of the challenges associated with working on big data projects, understanding the significance of ensuring data FAIRness and receiving guidance on achieving FAIRness with their own data. Furthermore, a key objective during the meeting was to discuss procedures for data exploitation in the context of big data, including handling and managing bad data. See Annex 9.4 for the participants list. The agenda of the workshop is listed below.

- Presentation by C. Herrera (ILL): *Introduction on big data and impact on (your) science.*
- Presentation by V. Favre-Nicolin (ESRF, head of algorithms and scientific data analysis group): *From big to FAIR data*. An overview of how LSFs (such as ESRF) handle and process big data within the FAIRness framework.
- Presentation by J. Flowers & S. Fuchs (KIT) from WP10: *Lessons in data management and automated data analysis.* An overview of how laboratories like KIT manage big data and perform automated data analysis from spectroscopic characterization techniques.

- Presentation and tutorial by S. Clark (SINTEF), WP7 lead: *Linked data principles and practical applications.*
    - Practical strategies to make your data FAIR.
    - Introductory demo: methodology for ontologized metadata using a specific case study.
    - Practical tools for creating and exploring ontologized metadata for your data.
    - Walk through of a hands-on example of creating, posting, and re-using data/metadata.
- All: WP5 discussion centered around enhancing pipelines and workflows to optimize data processing and analysis, specifically addressing exploitation procedures for big data/bad data handling and management. An important takeaway from the discussion was that scientists could greatly benefit from centralized online access to a repository of examples, tips and tutorials on how to handle and analysis big data.

7.1 Best practices to handle big data and bad data

Discussions during the data analysis workshop were highly productive and insightful and, from these we defined best practices for handling both big data and bad data in WP5. Two essential approaches need to be considered: data organization and data processing. In the following, we elaborate on these approaches and how to effectively implement them.

7.1.1 Data organization

Data organization is a fundamental task in general data analysis, regardless of the data type and size. It is essential for understanding the data and thus work efficiently. Figure 22 displays the seven steps we followed to effectively organize and handle the data.

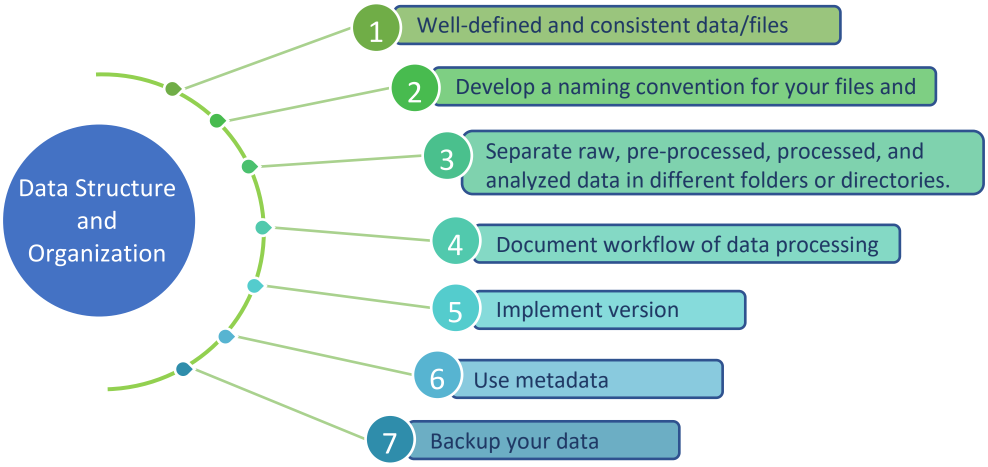

Data Structure and Organization
1 Well-defined and consistent data/files
2 Develop a naming convention for your files and
3 Separate raw, pre-processed, processed, and analyzed data in different folders or directories.
4 Document workflow of data processing
5 Implement version
6 Use metadata
7 Backup your data

Figure 22. Scheme illustrating best practices for data organization and handling.

Firstly, maintain a well-defined and consistent data and file structure: consider using a hierarchical folder structure to categorize and store different types of data. Secondly, develop a naming convention for files and folders, ensuring consistency and descriptiveness, and use acquisition dates for experimental data folders. Next, keep raw, pre-processed, processed and analyzed data in separate folders to track the data evolution and prevent any confusion or accidental overwriting. Additionally, incorporate metadata to your data, providing information about sample details and pre-processing steps to aid data discovery and understanding. A further step can be considered by ontologizing metadata following ontologies provided in WP7. Also, documenting workflows by keeping record of the steps and procedures followed during data processing is key to reproduce results and troubleshoot any issues that may arise. Consider also implementing version control using software like Git to track versions and facilitate collaboration with colleagues. And lastly, regular data backups are advised to prevent data loss using institutional servers, hard drives or cloud-based services. It may be helpful to set automatic synchronization routines for specific directories, e.g., by using `rsync -avz /path/to/local/folder/ username@remote_server:/path/to/remote/folder/` on Unix systems. Scheduling systematic synchronization tasks can be also implemented in Unix-like operating system with `cron` jobs.

In Section 7.1.3, we introduce a collaborative webpage where examples on data organization can be found.

7.1.2 Data processing

The way of processing and analyzing data has to be adjusted to the type of data. Here, we discuss three aspects for efficiently handling both big data and poor-quality (bad) data.

a. Scripting and optimizing scripts

In the context of big data analysis, scripting using common languages such as Python (which comes with many helpful libraries) is indispensable. When dealing with numerous datasets, a scripted algorithm can automatically be applied to all datasets, saving time and ensuring analysis consistency. Modifications to these algorithms can be easily done, making it efficient to adapt to changing data requirements. Every step in the process is documented, creating a strong framework that can be easily used with different datasets. In the following, we discuss three optimization aspects, for which Section 7.1.3 provides information on examples illustrating these approaches.

i. **Parallel processing for uncorrelated data**

One key strategy to optimize scripts is to use parallel processing, which consists in breaking down large datasets into smaller, more manageable batches that can be dealt with at the same time using the same algorithms. This approach significantly enhances processing speed and overall efficiency, allowing for quicker insights into the data. Parallel processing is most effective when applied to independent or loosely correlated datasets. Indeed, interdependencies between data elements require coordination and synchronization among the processing units which can potentially negate the performance gains that parallel processing typically provides. There exists a range of libraries and tools for parallel processing, providing the necessary infrastructure to distribute tasks across multiple processors or even clusters of machines, maximizing computational power and reducing processing times. One of the most used in Python is `multiprocessing`.

ii. **Object-Oriented Programming (OOP)**

The main idea in OOP is to organize the code by creating *objects*, that may represent different data structures or processing steps, which can have different *attributes* (data) and *methods* (functions). For instance, one can create objects to do specific actions such as load data, clean data or run analysis. Codes are then organized, easy to understand, and can be easily reusable.

iii. **Identifying bad data**

Identifying and handling bad data is a crucial consideration when dealing with big data. This identification depends on the specific technique and data type in use. We first need to define what we consider as bad data for any specific technique. For instance, raw neutron imaging and tomographic techniques often contain a certain number of white pixels, which may be due to scintillator/detector aging, cosmic ray interference, etc. Data visualization is one approach to identify problematic data, but it can be impractical when dealing with a large volume of data points. Statistical methods come to the forefront in such scenarios. In the case of imaging and tomography, software packages employ various algorithms to effectively clean data from artifacts and reduce noise.

b. Analytic techniques

Various analytic techniques, including statistical analysis, data mining, and machine learning, can be used to extract valuable insights and patterns from big data. Here, we provide some examples of these three approaches to analyze big data.

i. Statistical analysis

Statistical models offer numerous avenues for extracting meaningful information from big data sets. Descriptive statistics, such as means, medians, variances, and standard deviations, provide a comprehensive overview of the data characteristics. Basic correlation analysis, both linear and nonlinear, using parameters such as the Pearson correlation coefficient, helps uncover relationships within the data. Furthermore, techniques like Principal Component Analysis (PCA) are used to reduce datasets complexity by lowering dimensionality. Bayesian models are also useful for estimating probabilities and assessing uncertainty. The Kolmorogov-Smirnov test proves useful in determining whether two datasets come from the same underlying distribution.

ii. Data mining

Data mining refers to the discovery of patterns or associations within large datasets, with the goal of extracting useful information for decision-making or anomaly detection. This is done by using techniques such as clustering, which groups similar data points (e.g., using K-Means models), discovering unknown relationships between variables, finding unusual patterns, among others.

iii. Machine Learning

Machine learning offers a wide range of possibilities for leveraging data efficiently, representing a rapidly evolving field. The core concept involves developing algorithms and models capable of learning from existing data to make predictions. Models can be based on supervised learning (training data is labeled), unsupervised learning (e.g., clustering methods) and reinforcement learning. Machine learning models find application in image recognition (e.g., Section 5 for imaging segmentation), classification of spectral data (e.g., Section 2.4), and more. A significant challenge in this domain lies in the creation of large databases essential for training these models. Building robust training datasets is crucial to the success of machine learning applications, enabling the models to make accurate predictions.

c. Data Integration

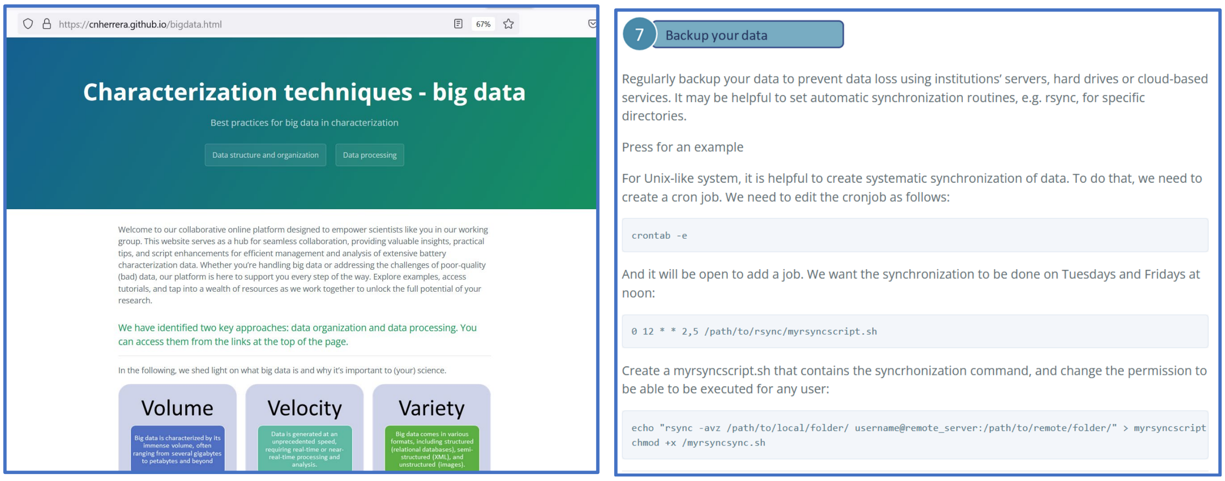
While sophisticated algorithms help uncover patterns, correlation, anomalies and predictions, data integration is a critical step in the workflow of data processing, especially when dealing with big data from diverse probes. It involves consolidating data from diverse sources into a unified format, allowing comprehensive analysis and correlation across different data types. This process includes identifying relevant sources and formats, transforming and cleaning data for consistency, and ensuring data quality. Having a well-defined data organization and structure ensures uniformity.

Also, metadata should be comprehensive, documenting data origin and transformation processes. Data should as well be easily accessed, allowing for smooth collaboration.

7.1.3 Collaborative webpage

Handling and analyzing data are key in scientific research. Often, scientists tend to focus on their own methodologies and data analysis techniques, and many are not specially trained to tackle the significant challenges that big data analysis can present. However, openly collaborating on methodologies for data handling and analysis with peers from diverse technical and scientific background is beneficial for everyone. With this vision in mind, we have created a collaborative webpage where scientists can explore best practices and guidelines related to data organization and processing (listed in this section), with access to examples and/or tutorials, and have the opportunity to start new topics and ask questions. Figure 23 presents screenshots from the web page.

The webpage is hosted on GitHub Pages, thus encouraging active participation. GitHub link: https://cnherrera.github.io/bigdata.html



https://cnherrera.github.io/bigdata.html

Characterization techniques - big data

Best practices for big data in characterization

Data structure and organization | Data processing

Welcome to our collaborative online platform designed to empower scientists like you in our working group. This website serves as a hub for seamless collaboration, providing valuable insights, practical tips, and script enhancements for efficient management and analysis of extensive battery characterization data. Whether you're handling big data or addressing the challenges of poor-quality (bad) data, our platform is here to support you every step of the way. Explore examples, access tutorials, and tap into a wealth of resources as we work together to unlock the full potential of your research.

We have identified two key approaches: data organization and data processing. You can access them from the links at the top of the page.

In the following, we shed light on what big data is and why it's important to (your) science.

Volume: Big data is characterized by its immense volume, often ranging from several gigabytes to petabytes and beyond

Velocity: Data is generated at an unprecedented speed, requiring real-time or near-real-time processing and analysis.

Variety: Big data comes in various formats, including structured (relational databases), semi-structured (XML), and unstructured (images).

7 Backup your data

Regularly backup your data to prevent data loss using institutions' servers, hard drives or cloud-based services. It may be helpful to set automatic synchronization routines, e.g. rsync, for specific directories.

Press for an example

For Unix-like system, it is helpful to create systematic synchronization of data. To do that, we need to create a cron job. We need to edit the cronjob as follows:

```
crontab -e
```

And it will be open to add a job. We want the synchronization to be done on Tuesdays and Fridays at noon:

```
0 12 * * 2,5 /path/to/rsync/myrsyncscript.sh
```

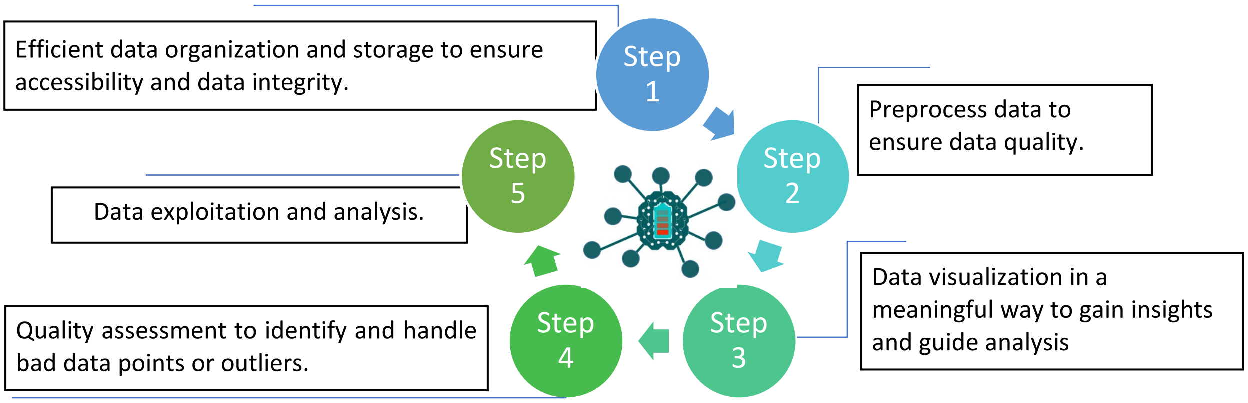
Create a myrsyncscript.sh that contains the syncrhonization command, and change the permission to be able to be executed for any user:

```
echo "rsync -avz /path/to/local/folder/ username@remote_server:/path/to/remote/folder/" > myrsyncscript
chmod +x /myrsyncsync.sh
```

Figure 23. Visual overview from the collaborative web page, showcasing examples, practical tips and best practices.

7.2 Data exploitation procedure for big data/bad data handling

Following our discussion on the data exploitation procedure for handling big data in the context of battery characterization data, we have outlined an approach for efficient analysis.


Efficient data organization and storage to ensure accessibility and data integrity.
Step 1
Step 2
Preprocess data to ensure data quality.
Step 3
Data visualization in a meaningful way to gain insights and guide analysis
Step 4
Quality assessment to identify and handle bad data points or outliers.
Step 5
Data exploitation and analysis.

In the first step, the use of a robust data management system, handled by WP9 with the BIG-MAP archive, to efficiently organize and store data, focusing on ensuring accessibility and data integrity. A collective effort is asked to all characterization researchers to upload data in a standardized BIG-MAP format, together with metadata for interoperability. Next, data preprocessing is essential to ensure data quality, including noise removal, artifact correction, and data normalization. In step three, the data exploration and visualization phase must consider the technique-dependent nature of data to achieve meaningful visualization to gain insights and to prepare subsequent analyses. Step four focuses on quality assessments, especially to address bad data, combining statistical analysis, data profiling, and technique-specific checks to pinpoint bad data points or outliers. Thresholds are defined to identify bad data and implement methods such as data imputation, filtering, or removal to handle this data. Finally, once data quality is ensured, we proceed to data exploitation and analysis. This encompasses a wide spectrum of analytical techniques, including statistical modeling, machine learning, and advanced data mining or data-driven modeling approaches.

In summary, this data exploitation procedure is designed to manage and analyze battery characterization data, emphasizing data quality, visualization, and a range of analytical techniques.

7.3 Ontologizing metadata

As crucial as the data itself, metadata plays a pivotal role in data analysis. Metadata provides the context necessary for a comprehensive and universal understanding of the data. In the context of FAIR principles for experimental data, many raw datasets already contain fundamental metadata related to the experimental acquisition. However, these metadata often lack critical details required for a comprehensive understanding of the experimental conditions and the probed sample. One significant challenge lies in the fact that different techniques may use varying terminology to describe the same aspects or employ identical keywords to define different attributes. This discrepancy presents a significant obstacle when attempting to establish connections between different datasets and having homogeneous datasets. Hence, the incorporation of ontologies into metadata becomes imperative to facilitate effective data linkage and comprehension.

Efforts to standardize metadata in materials characterization are ongoing. Under the OYSTER project, Romanos et. al. [34] have proposed CHADA, a new documentation structure for characterization data. It establishes protocols for multi-technique, multi-scale materials characterization and tools for data sharing, ontologies and standardized data documentation. CHADA describes four key concepts in the characterization workflow, (1) use case, with information of the testing environment, e.g., *sample;* (2) experiment, including process by which the methodological workflow is defined: *probe*, *detector*, *signal*, *noise*, etc.; (3) raw data, defined as a set of data provided directly as output by the experiment, typically in function of *time*, *position*, *energy*, etc.; and (4) data processing, identifying sequence of processes by which the raw data are transformed to obtain the final objective: the characterization properties.

While CHADA is easily interpretable by humans, the data is not structured for efficient and automated information retrieval. Based on CHADA, a new ontology named CHaracterization MEthodology Ontology (CHAMEO[15]) has been developed [35]. It provides a common framework for the development of ontologies related to specific techniques, enabling structured data that is machine-readable and facilitating semantic FAIR exchange of information. The CHAMEO ontology

[15] https://github.com/emmo-repo/domain-characterisation-methodology

models common aspects shared across different characterization techniques. For more detailed and technique-specific aspects, specialized ontologies should be developed by extending the CHAMEO definitions.

In the context of batteries, BattINFO (Battery INterface Ontology), developed within BIG-MAP by WP7, serves as a formal, machine-readable model encapsulating knowledge about electrochemistry and batteries. This ontology describes cycling processes, as well as battery definitions such as types, components, materials, and more.

While significant progress has been made in the development of ontologies, there is still work to be done to help and motivate experimental battery scientists to fully exploit ontologies for all characterization data. Within WP5, we have considered this and hosted, during the data analysis workshop, an ontology session led by S. Clark (WP7). WP5 scientists were introduced to the linked-data concept and followed an ontologization tutorial. WP5 scientists are committed to this comprehensive vision of linked data and are ready to use tools to ontologize their data.

Work is underway by WP7 in that regard. They have developed Batt-o-Matic, a prototype BIG-MAP app designed to display metadata from datasets, add new metadata entries, visualize through a knowledge graph, and generate ontologized metadata. The specification process must be carried out technique by technique, ensuring a comprehensive and consistent approach for the future.

8 Conclusions and perspectives

One main objective of WP5, and task 5.4, is to ensure that the data generated within WP5 remains comprehensible, reproducible, and interoperable. Through our data analysis work, we have witnessed that the standardization and automation of data analysis are essential components to allow reproducibility and interoperability, as well as to handle large amounts of data. Through diverse examples and a dedicated focus on specific techniques, we have highlighted the necessity of data analysis standardization and automation in the context of advanced material characterization.

Our work has spanned a range of scientific objectives, from the crystallographic study of LNO, with an emphasis on the analysis of XRD diffraction data, to our focus on vibrational spectroscopy encompassing Raman and FTIR techniques, the post-mortem analysis of reproducibility through the XPS technique and the advanced tomography segmentation employing machine learning methods. In response, we have developed and implemented tailored data analysis methods, common protocols and formats, tools, and algorithms, which have facilitated a streamlined and consistent data interpretation process across diverse experiments. Specifically, active learning for tomography segmentation and the on-the-fly LNO phase identification represent valuable assets for the research community, highlighting the potential of artificial intelligence in material characterization. By adopting standardization and automation, we are not only simplifying data analysis but also establishing a robust and reliable foundation for further advancements in battery research.

Inter-WP collaboration has been instrumental in our progress, as we have actively engaged in scientific multi-partner collaborations, and the development of insightful workshops to address specific challenges in data analysis such as the round-robin analysis, aiming at converging on XPS data analysis standardization protocols. Intra-WP collaboration has also played a pivotal role in the creation of automated tools for data analysis, exemplified by PRISMA (WP9) and the segmentation tool (WP11), both developed with the vision of meeting the needs of the characterization data analysis.

Looking ahead, the groundwork conducted within WP5 in data analysis lays the path for future work and opens up exciting possibilities. As we continue to accumulate large and diverse datasets, the need for even more advanced data analysis techniques and tools becomes evident. Multi-technique correlation based on data from the 5.3 experimental workflow, although outside the scope of the current task, presents an exciting challenge that we have just started addressing and will continue in future projects. Moreover, after being introduced to ontologies in the data workshop, we are committed to produce interoperable, linked data, enhancing the accessibility and utility of data for the research community.

9 Annexes

9.1 Description of experimental techniques and data sets

The different characterization techniques used in WP5 were outlined in the experimental matrix (D5.1). In this report, our attention is directed towards a specific subset of these techniques, which is detailed below. For each of them, essential information is provided, including a concise description of the technique itself, description of the raw data format, the typical software used for data analysis, and the main observables.

- **Raman**: Raman is a spectroscopic technique used to determine the vibrational modes of molecules based on the inelastic scattering of light. It provides a structural fingerprint by which molecules can be identified. When a material is irradiated with a monochromatic light, a small portion of the scattered light undergoes a change in frequency due to interactions with molecular vibrations, rotations, and other excitations within the material. This shift in frequency, known as the Raman shift, provides valuable information about molecular composition, structure, and bonding in the material.
    - Partner: ex situ and operando experiments by CTH.
    - Raw data: txt format
    - Pre-processed data: Raman shift vs. Intensity
    - Data analysis: PRISMA software, output data in txt.
    - Observables: electrolyte composition, chemical gradients (e.g., electrolyte concentration gradients near the electrodes), carbon structure and intercalation processes in anodes.

- **Fourier Transform Infrared (FTIR)**: FTIR is a spectroscopic technique used to determine the vibrational modes of molecules based on the absorption of IR radiation by the material. It is complementary to the Raman technique, probing vibrational excitations of the sample and thus providing a unique structural fingerprint for molecule identification. Selection rules for Raman and IR active modes are defined by the molecular vibrational symmetries. The vibration is Raman active if the vibrational mode results in the change of net polarizability, while the vibrational mode is IR active if there is a change in the net dipole moment. Therefore, the two techniques normally have different active modes and are complementary to each other.
    - Partner: ex situ and operando experiments by CTH and CSIC.
    - Raw data: txt format.
    - Pre-processed data: wavenumber vs. Intensity
    - Data analysis: PRISMA software, output data in txt.
    - Observables: electrolyte solvation structure and speciation, electrode/electrolyte interfaces.

- **XRD**: X-ray diffraction is an elastic scattering technique used to analyze the crystal structure of materials. A sample is irradiated with monochromatic X-rays and a diffraction (scattered) pattern is measured. The intensity and angles of the scattered X-rays are used to obtain information about the arrangement of atoms within the material structure, including crystal phases, lattice parameters, and crystallographic orientation. XRD is valuable for identifying and characterizing crystalline materials, determining their composition, and studying phase transitions. At synchrotron facilities, sensitivity is improved, thus giving higher spatial and temporal resolutions.

- Partner: ex situ and operando experiments by DTU and CEA.
- Raw data: depending on the instrument, it can be text file *.xy, .txt* or binary file *.raw*. Processed by data conversion or data reduction (mainly for synchrotron XRD data).
- Pre-processed data: 2θ angles (scattering angles) *vs.* intensity values (arbitrary units).
- Data analysis: Topas, FullProf, and PRISMA.
- Main observables: phases, lattice parameters, atomic coordinates, site occupancies.

- **XPS**: X-ray photoelectron spectroscopy employs the photoelectric effect to determine surface elemental composition. Penetration depth (10-100 µm) depends on the radiation source. X-ray beam hits the surface of a material. The energy absorbed that is needed to cause the core electron to be emitted and subsequently detected is unique to each element, allowing the use of binding energy (BE) to identify the elements present on the surface of the material. BE changes with the chemical environment of the element causing an effect known as *chemical shift*. Thus, XPS can estimate the thickness, uniformity, and surface chemistry of the sample's surface, being of great interest in SEI studies. XPS in synchrotron sources are called HAXPES, Hard X-ray Photoelectron Spectroscopy. The increased energy of the X-rays allows for deeper penetration into the sample, enabling the study of bulk properties and buried interfaces.
  - Partners: ex situ and operando experiments by CEA, UU and SOLEIL.
  - Raw data: format is in .spe.
  - Pre-processed data: Binding energy *vs.* intensity.
  - Data analysis: CasaXPS.
  - Main observables: SEI/CEI chemistry and composition

- **Electrochemistry:** experiments encompass measurements and analyses of electrical properties and behavior within battery systems. It provides insights into the performance, efficiency, and durability of batteries. Most of our partners use EC-Lab. Electrochemical data is integrated with other characterization data during operando experiments.
  - Partners: CSIC, NIC for cycling experiments, plus every partner that performs operando experiments.
  - Raw data: format is in .mpr
  - Pre-processed data: EC-Lab software
  - Data analysis: any visualization software such as KaleidaGraph, Gnuplot, Origin.
  - Main observables: potential (V), current density (A), capacity (mAh), Specific capacity (mAh/g)

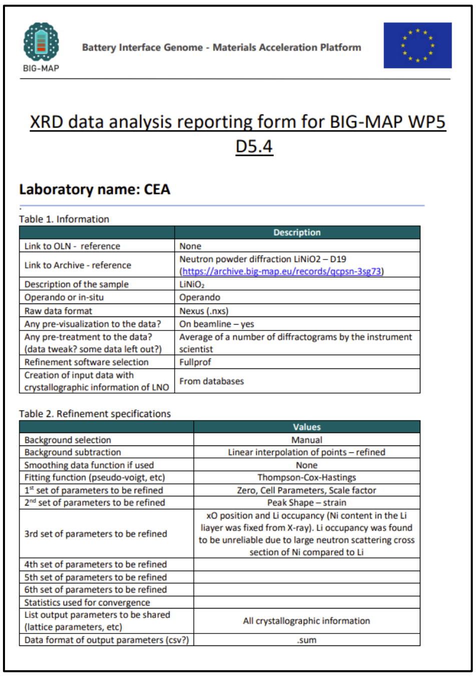
- **X-ray Tomography**: X-ray tomography makes use of an X-ray beam (either parallel or conical in a synchrotron set-up) to visualize in 3D a solid object including its surface and internal features. It is theoretically a non-destructive way of probing an object. The beam energy can vary widely depending on the set-up used and will influence directly the maximum size of the object that it is possible to acquire. Spatial resolution is dependent on a combination of factors including beam size, detector size and resolution, object position and size. The basic working principle is similar as to a CT medical scan, the object (or area to image) is fully illuminated by the X-ray beam and the transmitted beam is collected at numerous angles, here by rotating the sample, generating a set of images called radiographs or projections. These radiographs can then be used to reconstruct the 3D volume. Several algorithms and methods are available for the acquisition and reconstruction, from pure absorption contrast to phase retrieval, which will have an effect on the visible features inside the final reconstructed volume.

- Partners: CEA, ESRF and CTH at the ESRF and SOLEIL for operando and ex situ experiments.
- Raw data format: HDF5 format. Typical primary data is a count number (beam intensity) per pixel in the detector for every projection acquired.
- Pre-processed data: The radiographs can usually be previsualized with ImageJ/Fiji or Python if needed. The 3D volume reconstruction is usually performed through in-house developed software and facilities due to the process complexity and performance resource-intensiveness. The final volume usually consists in a set of tif images (either in separate files or in a single file stack).
- Data analysis can be performed through various software either commercial or open-source. Open-source software used in BIG-MAP are Fiji/ImageJ and Python.
- Main observables: 3D geometrical features like bulk morphology, phase volume fractions, porosity, interfacial area, particle size, fractures in the material, etc.

- **NPD**: Neutron powder diffraction. A monochromatic neutron beam impinges on the sample and the scattered neutrons are collected on a 2D detector. This technique, widely used in battery research, is employed to obtain the crystallographic structures of the material (i.e., the position of the atoms in the cell). By measuring the intensities and the positions of the scattered neutron spots/arcs in the resultant diffraction pattern, we can deduce the crystal structure, phase fractions, crystallite morphology, as well as strain information.
  - Partners: CEA at the ILL
  - Raw data format: NeXus format (HDF5).
  - Processed with Int3D or Mantid (at ILL).
  - Data analysis: refinement done with FullProf.
  - Main observables: lattice parameters, phases, atomic coordinates, site occupancies.

9.2 XRD standardization

In order to define a scheme for XRD data analysis standardization, specific partners performed parallel data analysis. Each of them filled out a report. The report includes pre-data analysis information such as description of the sample, experimental context, any requisites for data pre-processing or visualization and input crystallographic information. It also compiles details of the refinement process. This includes background selection and subtraction, consideration of data smoothing functions and selected fitting functions for the peak shaping. A categorical delineation of the sequence in which parameters were refined – first, second and so on and so forth — is listed, as well as the specific statistical measures used to achieve refinement convergence and their values. Figure 24 shows an example this report.



Battery Interface Genome - Materials Acceleration Platform

XRD data analysis reporting form for BIG-MAP WP5
D5.4

Laboratory name: CEA

Table 1. Information

	Description
Link to OLN - reference	None
Link to Archive - reference	Neutron powder diffraction LiNiO2 – D19 (https://archive.big-map.eu/records/qcpsn-3sg73)
Description of the sample	$LiNiO_2$
Operando or in-situ	Operando
Raw data format	Nexus (.nxs)
Any pre-visualization to the data?	On beamline – yes
Any pre-treatment to the data? (data tweak? some data left out?)	Average of a number of diffractograms by the instrument scientist
Refinement software selection	Fullprof
Creation of input data with crystallographic information of LNO	From databases

Table 2. Refinement specifications

	Values
Background selection	Manual
Background subtraction	Linear interpolation of points – refined
Smoothing data function if used	None
Fitting function (pseudo-voigt, etc)	Thompson-Cox-Hastings
1st set of parameters to be refined	Zero, Cell Parameters, Scale factor
2nd set of parameters to be refined	Peak Shape – strain
3rd set of parameters to be refined	xO position and Li occupancy (Ni content in the Li liayer was fixed from X-ray). Li occupancy was found to be unreliable due to large neutron scattering cross section of Ni compared to Li
4th set of parameters to be refined	
5th set of parameters to be refined	
6th set of parameters to be refined	
Statistics used for convergence	
List output parameters to be shared (lattice parameters, etc)	All crystallographic information
Data format of output parameters (csv?)	.sum

Figure 24. Example of report filled during XRD data analysis.

9.3 XPS data analysis workshop details

The XPS workshop was hosted at the EPN campus by ILL and ESRF, together with the coordination of CEA. It was part of a larger WP5 meeting that included another additional workshop (see Section 7) as well as a full day of scientific talks from all WP5 partners.

The workshop was guided by three XPS experts, A. Benayad, F. Capone and R. Fantin. They introduced the technique, presented a data analysis demo, tutored the participants and guided discussions.

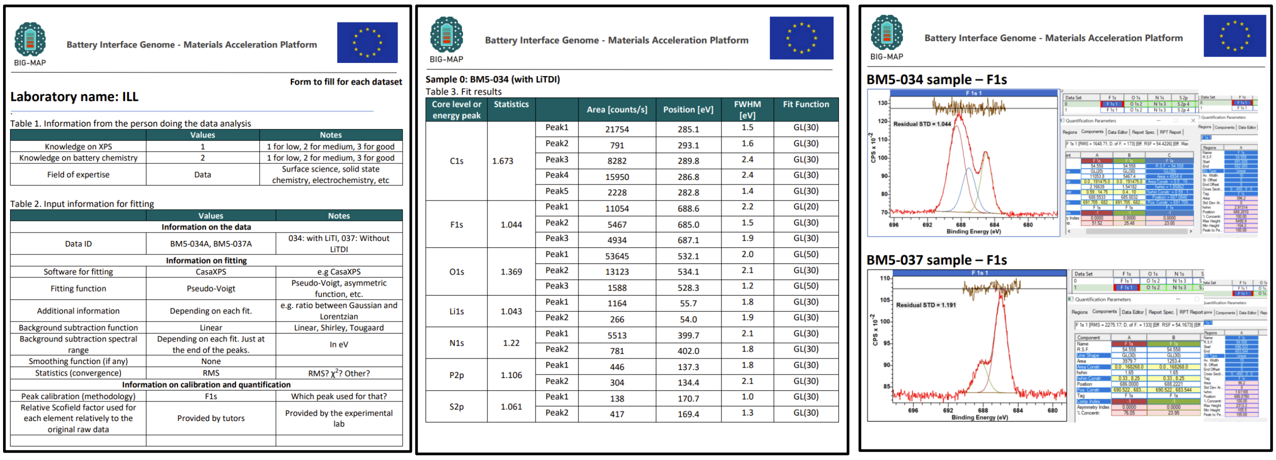
Tutors:

- Anass Benayad (CEA-Liten, WP5)
- Federico Capone (CNRS, WP5)
- Roberto Fantin (CEA-Liten, invited)

Participants:

- Francois CADIOU (ESRF, WP5)
- Jack FLOWERS (KIT, WP10)
- Stefan FUCHS (KIT, WP10)
- Cinthya HERRERA (ILL, WP5)
- Quentin JACQUET (CEA, WP5)
- Thanh Loan LAI (CEA-Liten, invited)
- Poul NORBY (DTU, WP5)
- Lucía PEREZ RAMIREZ (SOLEIL, WP5)
- Alexandre PONROUCH (CSIC-ICMAB, WP5)
- Shatakshi SAXENA (CEA-Liten, invited)
- Giorgio BARALDI (CID, WP6) – online
- Ajay GAUTAM (TUD, WP5) – online

During this exercise, the participants worked individually or in pairs on the data analysis. Supporting material such as handbooks, databases and key papers were available. Each participant filled out a report to document their analysis. Figure 25 displays an example of this report, which includes information of the expertise of the participant, the chosen parameters for the fitting, results of the fit for selected core levels, such as area, position, FWHM and fit function used, and examples of the fitting. Participants that had a CasaXPS license shared the CasaXPS project file, while those working with the demo version shared screenshots of the fitting.



Battery Interface Genome - Materials Acceleration Platform

Form to fill for each dataset

Laboratory name: ILL

Table 1. Information from the person doing the data analysis

	Values	Notes
Knowledge on XPS	1	1 for low, 2 for medium, 3 for good
Knowledge on battery chemistry	2	1 for low, 2 for medium, 3 for good
Field of expertise	Data	Surface science, solid state chemistry, electrochemistry, etc

Table 2. Input information for fitting

	Values	Notes
	Information on the data	
Data ID	BM5-034A, BM5-037A	034: with LiTI, 037: Without LiTDI
	Information on fitting	
Software for fitting	CasaXPS	e.g CasaXPS
Fitting function	Pseudo-Voigt	Pseudo-Voigt, asymmetric function, etc.
Additional information	Depending on each fit.	e.g. ratio between Gaussian and Lorentzian
Background subtraction function	Linear	Linear, Shirley, Tougaard
Background subtraction spectral range	Depending on each fit. Just at the end of the peaks.	In eV
Smoothing function (if any)	None	
Statistics (convergence)	RMS	RMS? χ^2? Other?
	Information on calibration and quantification	
Peak calibration (methodology)	F1s	Which peak used for that?
Relative Scofield factor used for each element relatively to the original raw data	Provided by tutors	Provided by the experimental lab

Battery Interface Genome - Materials Acceleration Platform

Sample 0: BM5-034 (with LiTDI)

Table 3. Fit results

Core level or energy peak	Statistics		Area [counts/s]	Position [eV]	FWHM [eV]	Fit Function
C1s	1.673	Peak1	21754	285.1	1.5	GL(30)
		Peak2	791	293.1	1.6	GL(30)
		Peak3	8282	289.8	2.4	GL(30)
		Peak4	15950	286.8	2.4	GL(30)
		Peak5	2228	282.8	1.4	GL(30)
F1s	1.044	Peak1	11054	688.6	2.2	GL(20)
		Peak2	5467	685.0	1.5	GL(30)
		Peak3	4934	687.1	1.9	GL(30)
O1s	1.369	Peak1	53645	532.1	2.0	GL(50)
		Peak2	13123	534.1	2.1	GL(30)
		Peak3	1588	528.3	1.2	GL(50)
Li1s	1.043	Peak1	1164	55.7	1.8	GL(30)
		Peak2	266	54.0	1.9	GL(30)
N1s	1.22	Peak1	5513	399.7	2.1	GL(30)
		Peak2	781	402.0	1.8	GL(30)
P2p	1.106	Peak1	446	137.3	1.8	GL(30)
		Peak2	304	134.4	2.1	GL(30)
S2p	1.061	Peak1	138	170.7	1.0	GL(30)
		Peak2	417	169.4	1.3	GL(30)

Battery Interface Genome - Materials Acceleration Platform

BM5-034 sample – F1s

BM5-037 sample – F1s

Figure 25. Example of the reports filled out by each of the participants.

9.4 Big data and FAIRness workshop participant list

Together with the XPS workshop, we also organized a big/bad data and FAIRness workshop, further described in Section 7. Here is a list of the participants.

Tutor:

- Simon CLARK (SINTEF, WP7)

Participants:

- Didier BLANCHARD (ESRF, WP5)
- Francois CADIOU (ESRF, WP5)
- Federico CAPONE (CNRS, WP5)
- Jack FLOWERS (KIT, WP10)
- Stefan FUCHS (KIT, WP10)
- Cinthya HERRERA (ILL, WP5)
- Quentin JACQUET (CEA, WP5)
- Sandrine LYONNARD (CEA, WP5)
- Lucía PEREZ RAMIREZ (SOLEIL, WP5)
- Giorgio BARALDI (CID WP6) – online
- Andy NAYLOR (UU, WP5) – online

10 References

[1] Coudert, F. X., 2017, Reproducible Research in Computational Chemistry of Materials, *Chem. Mater*., 29, 2615-2617.

[2] Alston, J. M. & Rick, J. A, 2021, A beginner's guide to conducting reproducible research, *Bull. Ecol. Soc. Am.,* 102, 2, 1801.

[3] Bollen, K., Cacioppo, J. T., Kaplan, *et al.*, 2015, Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*, Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.*

[4] Flores, E., Mozhzhukhina, N., Li, X. *et al.*, 2022, PRISMA: A Robust and Intuitive Tool for High-Throughput Processing of Chemical Spectra. *Chemistry–Methods*, 2, e202100094.

[5] Ohzuku, T., Ueda, A. & Nagoyama, M. *et al*., 1993, Electrochemistry and Structural Chemistry of $LiNiO_2$ (R3m) for 4 Volt Secondary Lithium Cells, *J. Electrochem. Soc.,* 140, 1862.

[6] Delmas, C., Ménétrier, M., Croguennec, L., *et al*., 1999, Lithium batteries: a new tool in solid state chemistry, *International Journal of Inorganic Materials*, 1, 1, 11-19.

[7] Bianchini, M., Roca-Ayats, M., Hartmann, P., *et al*., 2019, There and Back Again-The Journey of LiNiO2 as a Cathode Active Material, *Angew. Chem. Int. Ed. Engl*., 58, 10434–10458

[8] Li, W., Reimers, J. & Dahn, J. R., 1993, In situ x-ray diffraction and electrochemical studies of Li1−xNiO2, *Solid State Ionics*, 67, 123-130.

[9] Yoon, C. S., Jun, D.-W., Myung S.-T., Sun Y.-K., 2017, Structural Stability of LiNiO2 Cycled above 4.2 V, *ACS Energy Lett.*, 2, 1150-1155.

[10] Rietveld, H. M., 1969, A profile refinement method for nuclear and magnetic, *Journal of Applied Crystallography,* 2, 65.

[11] Rodríguez-Carvajal, 1993, Recent advances in magnetic structure determination by neutron powder diffraction, *Physica B*, 192, 55.

[12] Coelho, A. A., 2018, TOPAS and TOPAS-Academic: an optimization program integrating computer algebra and crystallographic objects written in C++, *J. Appl. Cryst*., 51, 210-21.

[13] McCusker, L. B., Von Dreele, R. B., Cox, *et al.*, 1999, Rietveld refinement guidelines*, J. Appl. Cryst.*, 32, 36-50.

[14] Eilers, P. H. C., 2003, A Perfect Smoother, *Anal Chem*, 75, 3631–3636.

[15] Eilers, P. H. C. & Boelens, H. F. M., 2005, Baseline correction with asymmetric least squares smoothing, *Leiden University Medical Centre Report*, 1, 5.

[16] Dedryvère, R., Leroy, S., Martinez, H., *et al.*, 2006, XPS Valence Characterization of Lithium Salts as a Tool to Study Electrode/Electrolyte Interfaces of Li-Ion Batteries, *J. Phys. Chem. B*, 110, 12986-12992.

[17] Maibach, J., Rizell, J., Matic, A. & Mozhzhukhina, N., 2023, Toward Operando Characterization of Interphases in Batteries, *ACS Materials Lett.*, 5, 9, 2431–2444.

[18] Dou, X., Hasa, I., Saurel, D., *et al.*, 2019, Hard carbons for sodium-ion batteries: Structure, analysis, sustainability, and electrochemistry, *Materials Today,* 23, 87–104.

[19] Marino, C., Cabanero, J., Povia, M. & Villevieille, C., 2018, Biowaste Lignin-Based Carbonaceous Materials as Anodes for Na-Ion Batteries, *J Electrochem Soc,* 165, A1400–A1408.

[20] Sadezky, A., Muckenhuber, H., Grothe, H., *et al.,* 2005, Raman microspectroscopy of soot and related carbonaceous materials: Spectral analysis and structural information, *Carbon N Y,* 43, 1731–1742.

[21] Pawlyta, M., Rouzaud, J.-N. & Duber, 2015, S. Raman microspectroscopy characterization of carbon blacks: Spectral analysis and structural information, *Carbon N Y,* 84, 479–490.

[22] Moon, H., Zarrabeitia, M., Frank, E., *et al.*, 2021, Assessing the Reactivity of Hard Carbon Anodes: Linking Material Properties with Electrochemical Response Upon Sodium- and Lithium-Ion Storage, *Batter Supercaps,* 4, 960–977.

[23] Demtröder, W., 2003, Widths and Profiles of Spectral Lines, *Laser Spectroscopy,* 59–95 (Springer).

[24] Váczi, T., 2014, A new, simple approximation for the deconvolution of instrumental broadening in spectroscopic band profiles, *Appl Spectrosc,* 68, 1274–1278.

[25] Virtanen, P., Gommers, R., Oliphant, T. E., *et al.*, 2020, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat Methods,* 17, 261–272.

[26] Maiti, M., Krishnamoorthy, A. N., Mabrouk, Y., *et al.*, 2023, Mechanistic understanding of the correlation between structure and dynamics of liquid carbonate electrolytes: impact of polarization, *Physical Chemistry Chemical Physics,* 25, 20350–20364.

[27] Aroca, R., Nazri, M., Nazri, G. A., *et al*., 2000, Vibrational Spectra and Ion-Pair Properties of Lithium Hexafluorophosphate in Ethylene Carbonate Based Mixed-Solvent Systems for Lithium Batteries, *J Solution Chem.,* 29, 1047–1060.

[28] Han, S.-D., Yun, S.-H., Borodin, O., *et al*., 2015, Solvate Structures and Computational/Spectroscopic Characterization of LiPF6 Electrolytes, *The Journal of Physical Chemistry C,* 119, 8492–8500.

[29] Seo, D. M., Reininger, S., Kutcher, M., *et al.*, 2015, Role of Mixed Solvation and Ion Pairing in the Solution Structure of Lithium Ion Battery Electrolytes, *The Journal of Physical Chemistry C*, 119, 14038–14046.

[30] Cresce, A. V, Russell, S. M., Borodin, O., *et al.,* 2017, Solvation behavior of carbonate-based electrolytes in sodium ion batteries, *Physical Chemistry Chemical Physics*, 19, 574–586.

[31] Ronneberger, O., Fischer, P., Brox, T., 2015, U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, *Springer, LNCS*, 9351, 234-241.

[32] Leonelli, S., 2020, Scientific Research and Big Data, *The Stanford Encyclopedia of Philosophy (Summer 2020 Edition),* Edward N. Zalta (ed.), URL: https://plato.stanford.edu/archives/sum2020/entries/science-big-data/

[33] Wilkinson, M. D., Dumontier, M., Aalbersberg, I., *et al.,* 2016, The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 3, 160018.

[34] Romanos, N., Kalogerini, M., Koumoulos, E. P., *et al*., 2019, Innovative data management in advanced characterization: implications for materials design*, Mater. Today Commun.*, 20, 100541.

[35] Del Nostro, P., Goldbeck, G., Toti, D., 2022, CHAMEO: An ontology for the harmonisation of materials characterisation methodologies, *Applied Ontology Volume*, 17, 32022, 401-421.