



KNOWSKITE-X

Knowledge-driven fine-tuning of perovskite-based electrode materials for reversible Chemicals-to-Power devices

Managing Data for Machine Learning: The KNOWSKITE-X Approach to a Shared and Adaptive Knowledge Base

Franz Martin Rohrhofer (Know Center Research GmbH, AUT)
frohrhofer@know-center.at

MatCHMaker workshop @ EMMC 2025
APR 7, 2025 | TU Wien, VIENNA



Funded by the
European Union

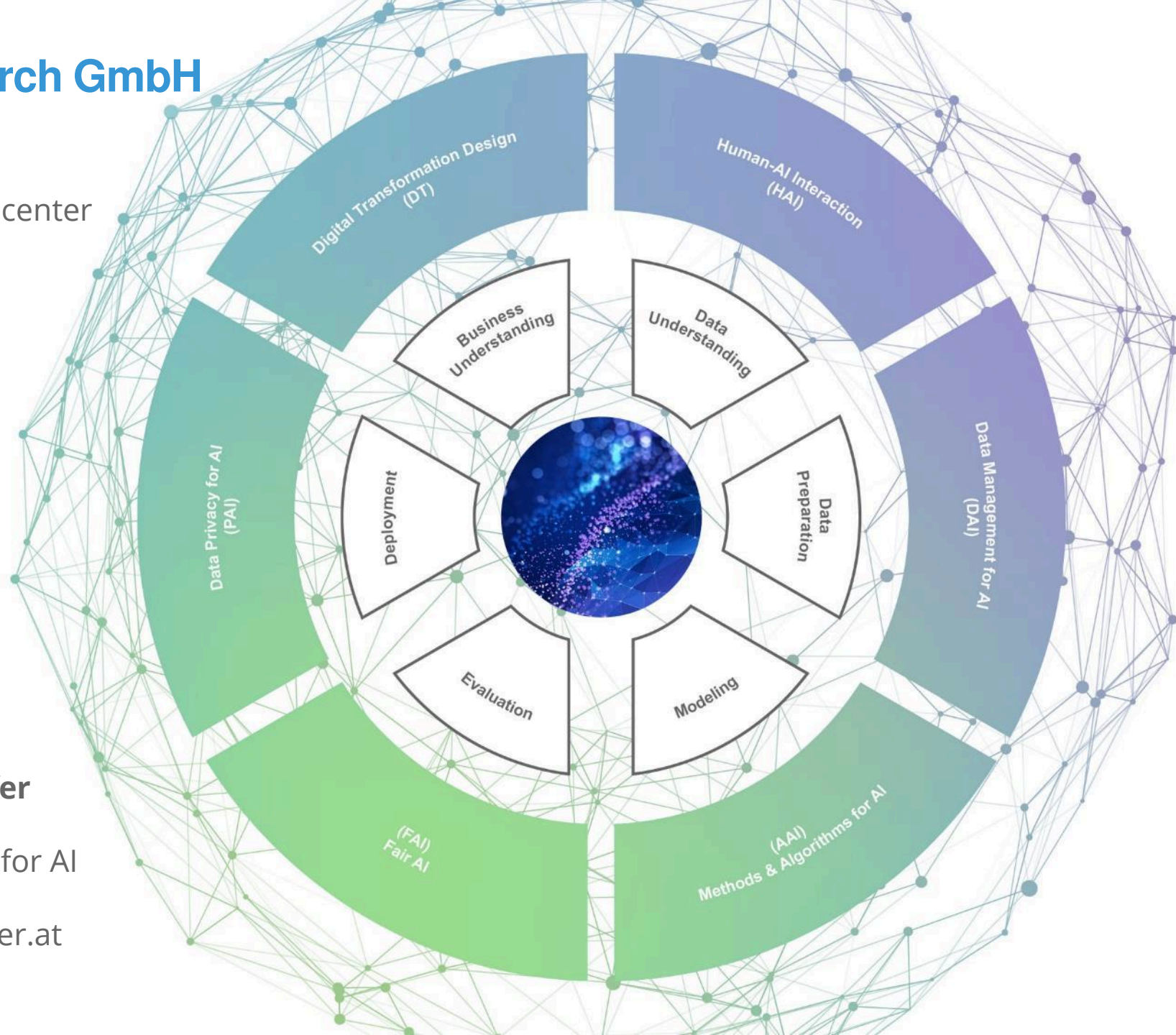
“Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.”

Leading European innovation and research center
for trusted AI and Data Science.



Franz Martin Rohrhofer
Senior Researcher
Methods & Algorithms for AI

frohrhofer@know-center.at





KNOWSKITE-X

Knowledge-driven fine-tuning of perovskite-based electrode materials for reversible Chemicals-to-Power devices

Grant Agreement Number	101091534
Project Full Title	Knowledge-driven fine-tuning of perovskite-based electrode materials for reversible Chemicals-to-Power devices
Project Acronym	KNOWSKITE-X
Topic	HORIZON-CL4-2022-RESILIENCE-01-19
Type of action	HORIZON-RIA
Granting authority	HADEA
Start date	01 January 2023
Duration	48 months
EU Contribution	5.168.000.00 Euro



Funded by the
European Union

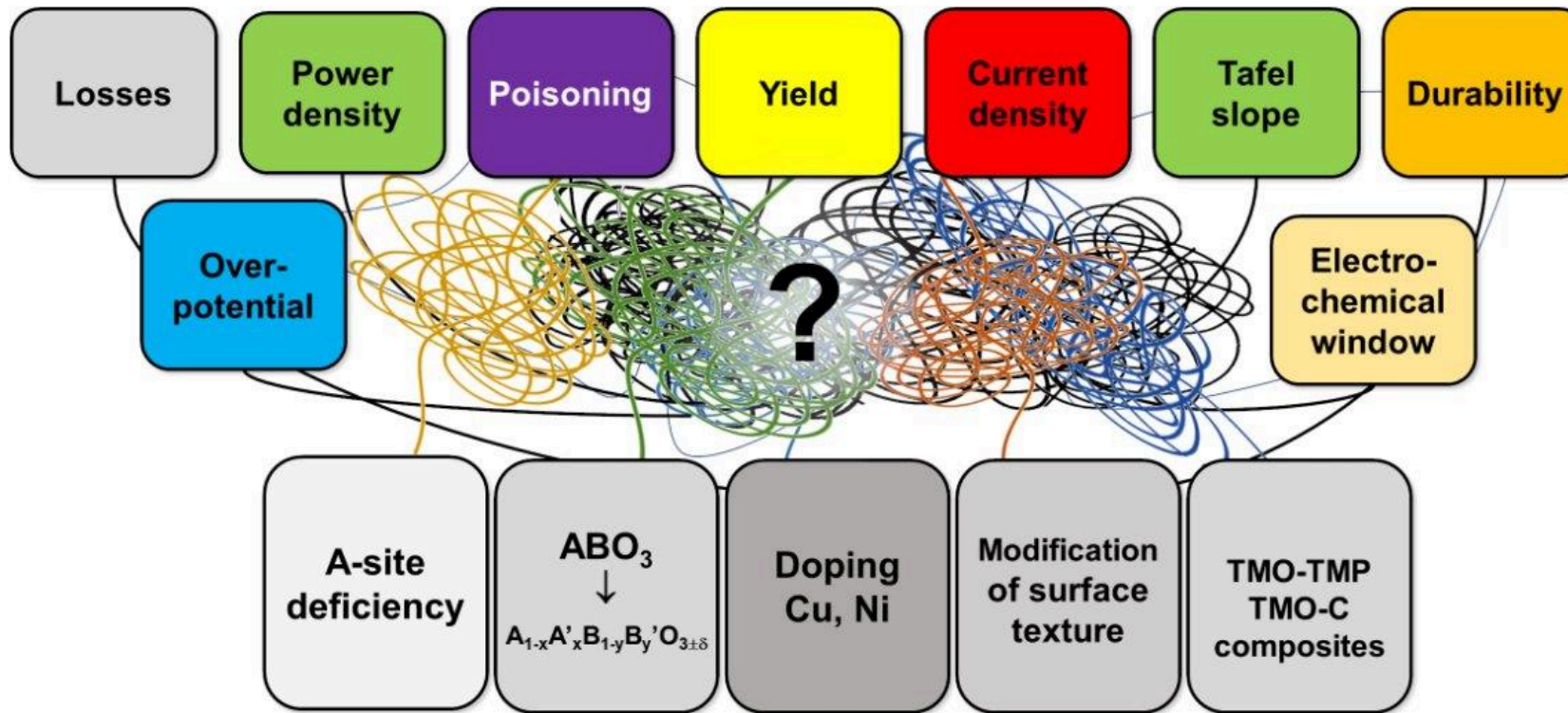
“Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.”

Project

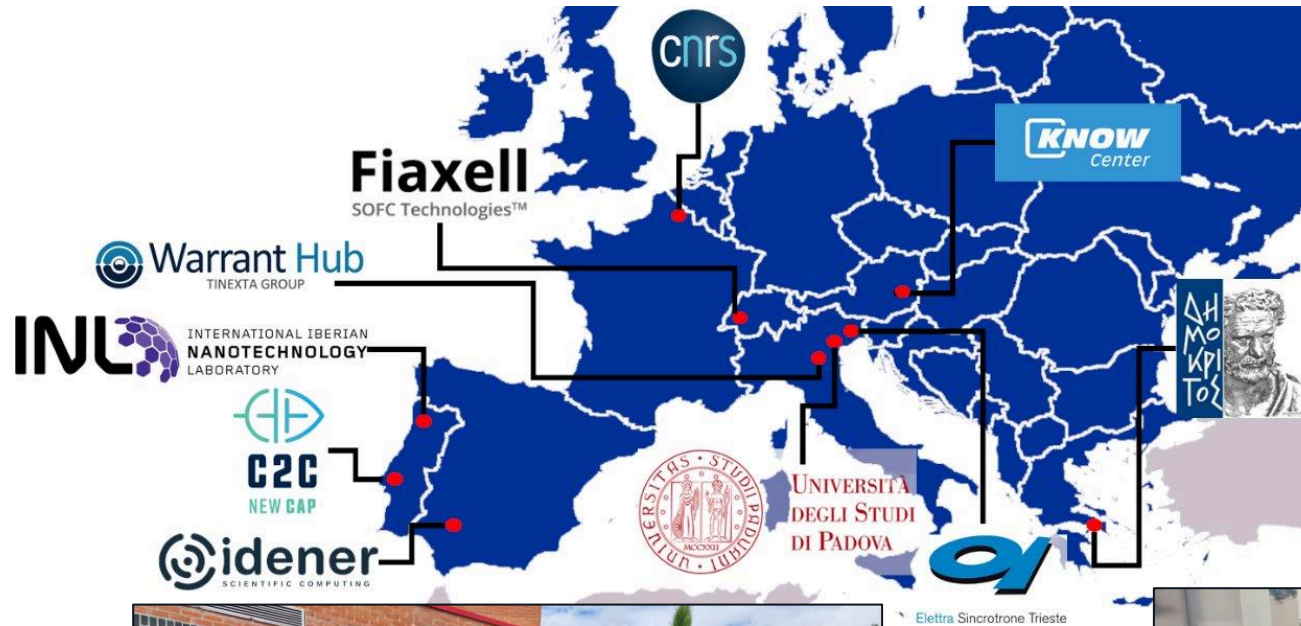
The project will demonstrate a science-based approach to the development of **electrode materials** forming key parts of **reversible chemical-to-power cells**. Such devices operate in two modes: in fuel cell (FC) mode, it converts hydrogen into electricity whereas when operating as electrolyser cell (EC), it uses excess electricity to form hydrogen from water electrolysis. This versatility enables the integration of intermittent renewable energy sources with the electrical grid by storing the excess energy as carbon-free chemical fuel. In particular, the project targets **mixed oxides with perovskite structure** with minimised critical content while keeping highest possible performances and targeting fair economic viability.

Objectives

The main objective of the KNOWSKITE-X project is to boost the development of materials for energy applications by combining state-of-art approaches together with the empowerment of knowledge discovery allowed by artificial intelligence (AI). In particular, the project integrates a smart combination of advanced technologies, involving tailor-made **materials preparation**, harmonised and ground-breaking **characterisation methods**, **multi-scale modelling** and **AI-enabled tools**. This corpus of open-minded, innovative, reliable, and use-relevant methodologies targets the discovery of the scientific knowledge required to sustain the rational design of optimized candidate electrode materials.

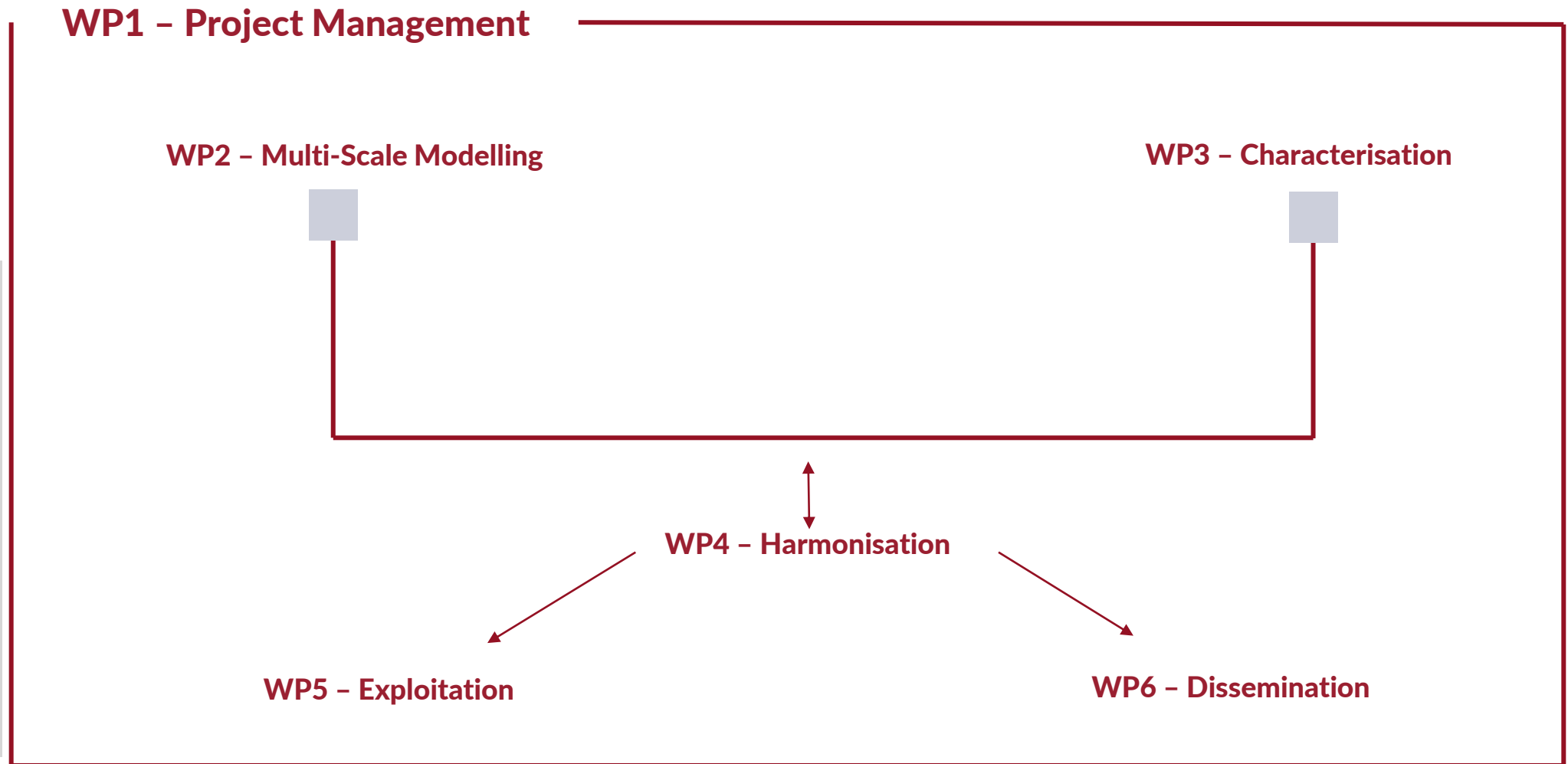


The Project Partners of KNOWSKITE-X



Funded by the European Union

“Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.”



➤ Interoperability of Data

- Started from the beginning of the project
- Identify **which data are targeted and how they will be managed**
- Develop an **ontology to organise the data and metadata** from experiments, modelling and data mining activities
- Define metadata standards, semantic, documentation, technological solutions, methodology, etc.

➤ Data Management Plan

- Key aspect in KNOWSKITE-X and is **iteratively updated**
- **Initial questionnaire** to collect information about the datasets (based on the template provided by the European Commission)
- Documentation about the handling of data during and after the end of the project
- Final purpose is the implementation of **FAIR datasets** (findable, accessible, interoperable, and reusable)

Table 2. Type of data potentially appearing in KNOWSKITE-X

Observational	Collected in real-time (field data, operator, laboratory notes, pictures, etc.) and obtained in the demo activities (WP2, WP3 and WP5)
Experimental	Collected during the testing of materials or prototypes and the characterisation and identification process (WP2 and WP3)).
Simulation	Models developed in WP2 will generate simulation data (prototypes, schemes, scripts, etc.). The data appearing from this activity will be replicable.
Derived	Datasets that already exist are used to develop reproducible derived data. Data obtained from experimental datasets may be used as inputs for modelling activities, environmental and techno-economic assessments, calculation and interpretation of data, graphics, analytics of data, results analysis, etc. The partnership additionally uses internal coding to alter data from their devices.
Reference/Canonical	A collection of published and/or curated (peer-reviewed) datasets, waste codes (EU/Basel), and data sets make up a reference or canonical data. These kinds of data will be utilised as input for modelling tasks (WP2) and sustainability assessments (WP5), along with experimental data.

Identifier	RAMAN spectra (ex situ and in situ) of divided oxides and thin films.
Dataset description	Raman intensity (Y, arbitr. unit) as a function of wavenumber (X, cm ⁻¹)
Purpose of the data	Structural investigation at molecular level, identification of minority phases, possibly in situ/operando.
Type of data	Raw spectra (XY), curated / processed XY data (e.g. baseline subtraction, fluorescence removal, normalisation...), Figures.
Form of the data	Origin files, Images (Figures), XY columns text
Format of the data	.txt (XY columns), .png or .eps (figures), .spc (series of spectra)
Origin of the data	Generated by our spectrometer
Data Utility	Scientific community (readers of publications), R&D engineers, chemists
Re-use of existing data	Re-use foreseen for the purpose of comparison, knowledge generalisation...
Data set is:	Growing, publishable, documented (metadata readme associated file), re-usable, FAIR-ready
Size of data	100 Mo of XY + metadata files Figures presenting curated data: ca. 1 Go

➤ Interoperability of Data

- Started from the beginning of the project
- Identify **which data are targeted and how they will be managed**
- Develop an **ontology to organise the data and metadata** from experiments, modelling and data mining activities
- Define metadata standards, semantic, documentation, technological solutions, methodology, etc.

➤ Data Management Plan

- Key aspect in KNOWSKITE-X and is **iteratively updated**
- **Initial questionnaire** to collect information about the datasets (based on the template provided by the European Commission)
- Documentation about the handling of data during and after the end of the project
- Final purpose is the implementation of **FAIR datasets** (findable, accessible, interoperable, and reusable)

Table 2. Type of data potentially appearing in KNOWSKITE-X

Observational	Collected in real-time (field data, operator, laboratory notes, pictures, etc.) and obtained in the demo activities (WP2, WP3 and WP5)
Experimental	Collected during the testing of materials or prototypes and the characterisation and identification process (WP2 and WP3).
Simulation	Models developed in WP2 will generate simulation data (prototypes, schemes, scripts, etc.). The data appearing from this activity will be replicable.
Derived	Datasets that already exist are used to develop reproducible derived data. Data obtained from experimental datasets may be used as inputs for modelling activities, environmental and techno-economic assessments, calculation and

Pain Points

- Many partners → **Diverse approaches**
- Many methods → **Inconsistent workflows**
- Much data → **Large, complex datasets**
- Many data formats → **Interoperability issues**
- Varying quality → **Uncertain reliability**

Result: Challenging data integration & analysis

Data Utility	Scientific community (readers of publications), R&D engineers, chemists
Re-use of existing data	Re-use foreseen for the purpose of comparison, knowledge generalisation...
Data set is:	Growing, publishable, documented (metadata readme associated file), re-usable, FAIR-ready
Size of data	100 Mo of XY + metadata files Figures presenting curated data: ca. 1 Go

➤ Dedicated and Private Online Repository

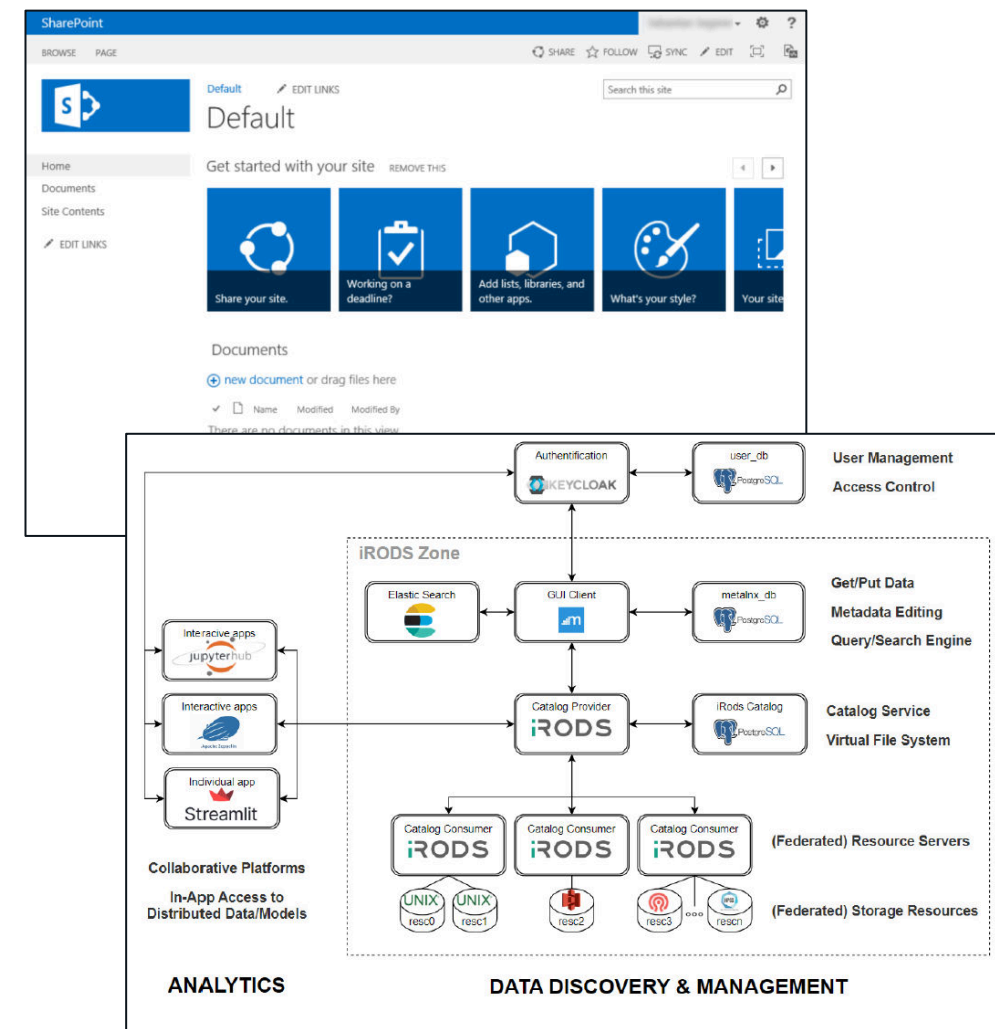
- **Microsoft SharePoint** hosted in WarrantHub server
- Internal communication channel
- Upload and assess documents, reports, and other kinds of files

➤ GitHub Repository

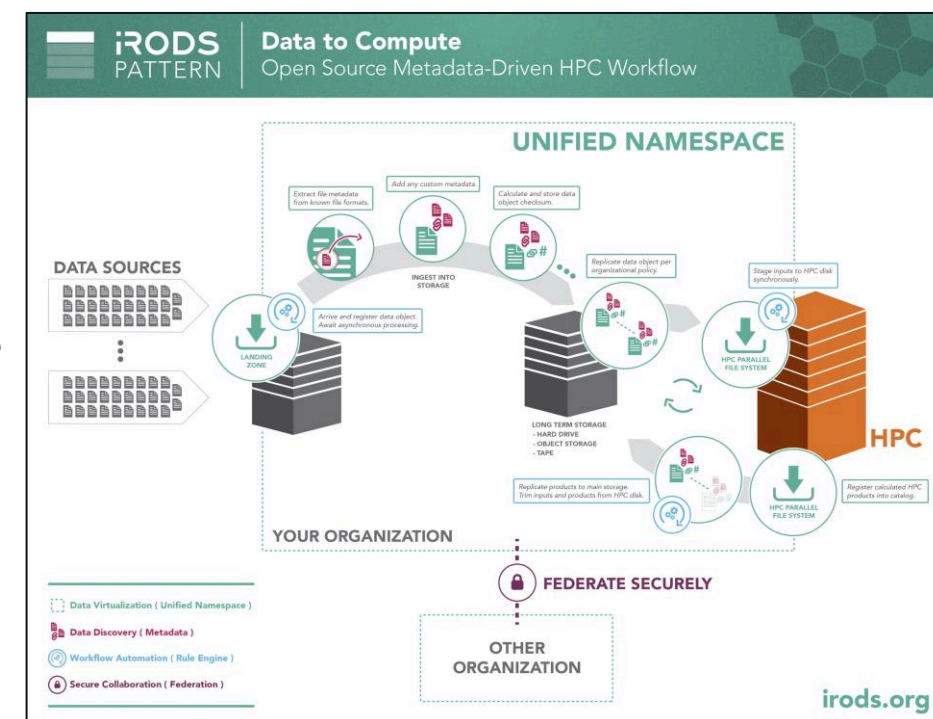
- Primarily used to **share code**, e.g. machine learning models
- Development of internal data management platform

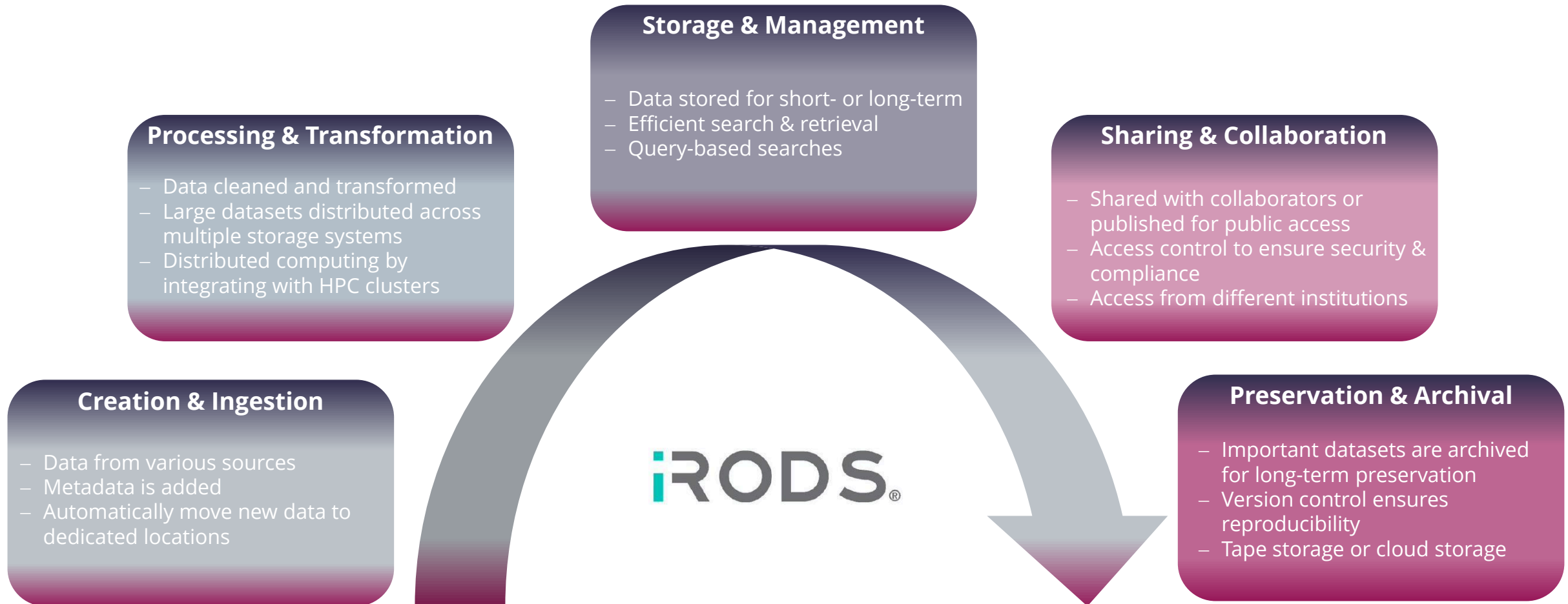
➤ Internal Data Management Platform

- Hosted as repository
- Is built on the **iRODS (integrated rule-oriented data system)** data management system
- Provides unified interface for accessing, managing and sharing data
- Provides **shared and adaptive interface** with various web-applications (e.g., Metalnx, streamlit, JupyterHub)
- Iteratively designed, adjusted, and improved based on the needs of project partners



- **Open-Source Data Management Software**
- **Policy-Driven Data Management (Automation & Rules Engine)**
 - Automates data workflows using built-in Rules Engine (e.g. pre-processing)
 - Defines metadata tagging, access control, and archival
- **Metadata-Rich File Management (FAIR Data Principles)**
 - Every file and collection can store custom metadata
 - Supports metadata-based queries (e.g., based on attributes and units)
- **Distributed & Federated Data Management**
 - Supports federation (e.g., connect and share data across institutions)
 - Enables hybrid data management through on-premises and cloud storage
- **Strong Access Control & Auditing**
 - Fine-grained access control at different levels (e.g., files or collections)
 - Auditing & logging enables tracking of access, modification, and deletion
- **Scalability for Large Scientific Workflows**
 - Designed for HPC, research data, and big data storage
 - Handles large-scale data collections with hierarchical organization





➤ Main Page / Landing Page (Streamlit)

- Serves as entry point to the user
- Provides links to other applications



Data Management (Metalnx)

- Provides **GUI for iRODS**
- Is used to
 - ✓ Easily upload and download files
 - ✓ Manage metadata
 - ✓ Manage groups and permissions



Workflow Editor (Streamlit)

- Used to **create & view MODA/CHADA workflows**
- Connects to iRODS to retrieve/identify files and folders
- Features templates to easily create new workflows



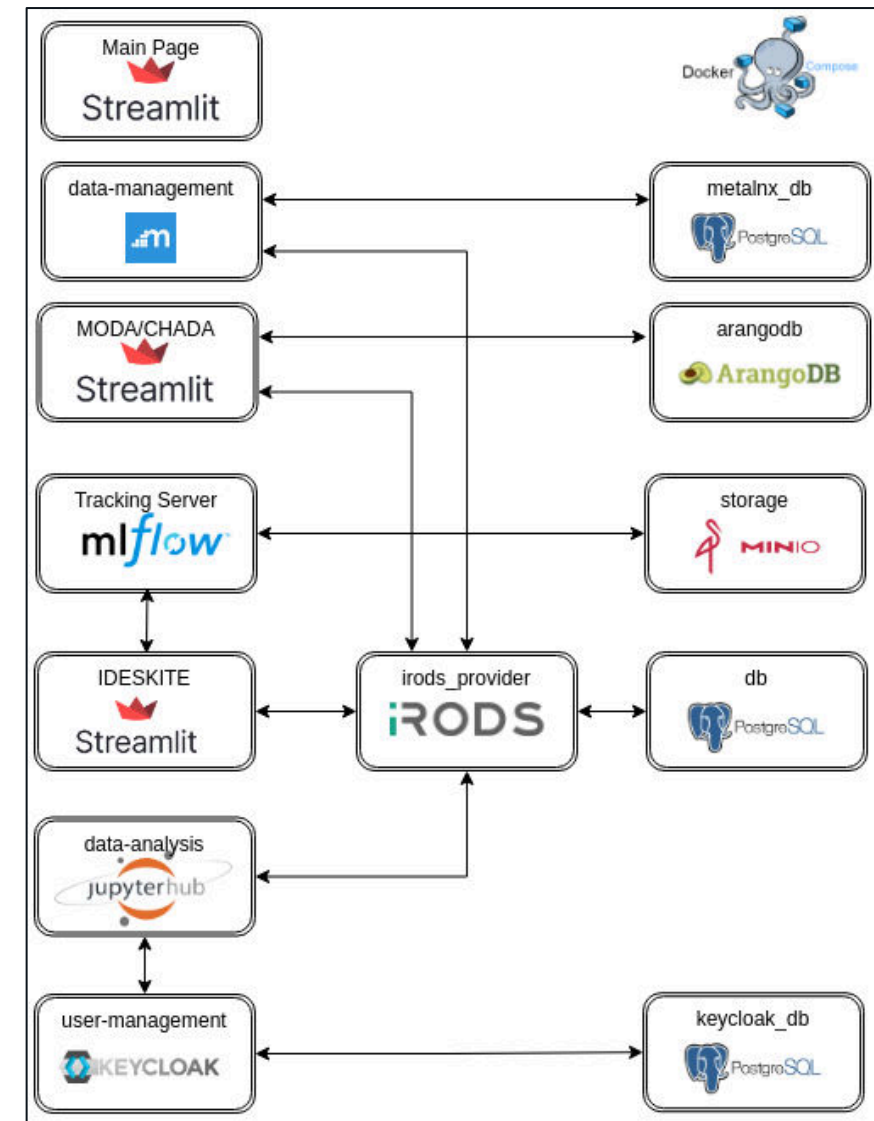
IDEskite (Streamlit)

- Provides a **machine learning platform** to train models
- Connects to iRODS to select training data



Data Analysis (JupyterHub)

- Provides analysis platform for data
- Connects to iRODS to retrieve data



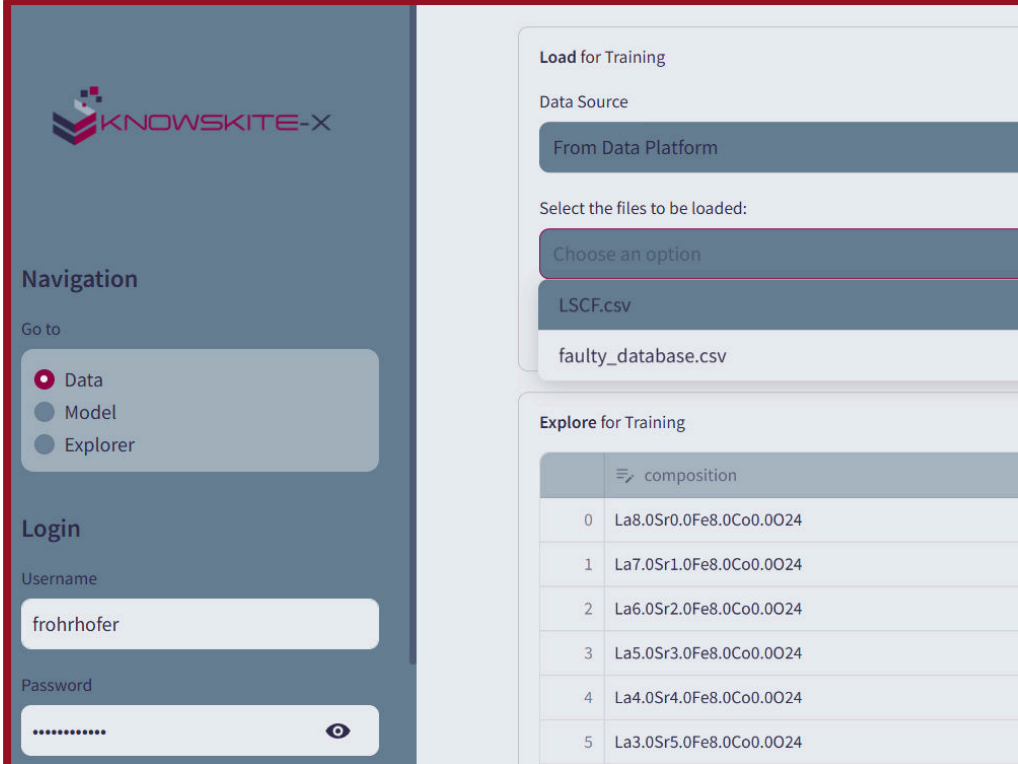
➤ Adaptivity through Python API

- iRODS can be easily interconnected with external frameworks, e.g. Metalnx, streamlit, JupyterHub
- Basically one line of code

```
self.session = iRODSSession(host=IRODS_HOST_NAME,  
                             port=IRODS_PORT, user=usr,  
                             password=pwd, zone=IRODS_ZONE)
```

➤ Adaptivity through Streamlit

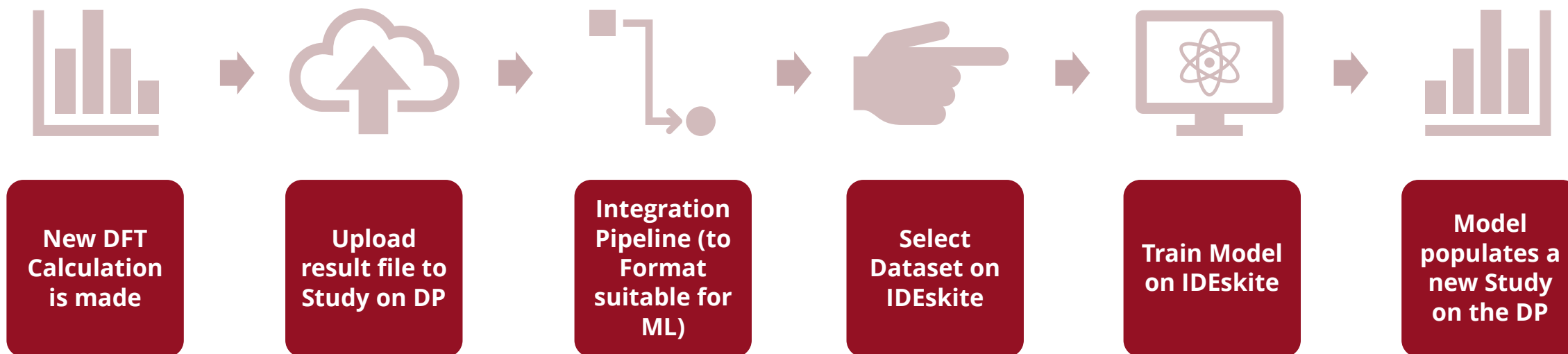
- *Streamlit*: open-source Python framework used to create
 - **interactive data applications,**
 - **dashboards,** and
 - **machine learning interfaces**
- No need for frontend development
- Automatically creates Widgets (sliders, buttons, etc...)
- Supports real-time update



The image shows a Streamlit web application for KNOWSKITE-X. It features a sidebar with a navigation menu (Data, Model, Explorer) and a login section with fields for Username (frohrhofer) and Password. The main content area has two sections: 'Load for Training' with a 'Data Source' dropdown set to 'From Data Platform', a file selection area showing 'LSCF.csv' and 'faulty_database.csv', and 'Explore for Training' which displays a table of material compositions.

	composition
0	La8.0Sr0.0Fe8.0Co0.0O24
1	La7.0Sr1.0Fe8.0Co0.0O24
2	La6.0Sr2.0Fe8.0Co0.0O24
3	La5.0Sr3.0Fe8.0Co0.0O24
4	La4.0Sr4.0Fe8.0Co0.0O24
5	La3.0Sr5.0Fe8.0Co0.0O24

Workflow Concept Example (DFT)



Hybrid Modeling of Mixed Perovskite Oxides with Gaussian Process Regression: Overcoming Data Scarcity

Franz M. Rohrhofer^{1,*}, Jean-François Paul², Bernhard C. Geiger¹, Elise Berrier²

¹) Know Center Research GmbH, Sandgasse 34, 8010 Graz, Austria
²) Université Lille, CNRS UMR8181, Unité de Catalyse et de Chimie du Solide, UCCS, F-59655 Villeneuve d'Ascq, France
^{*}) Corresponding author: frohrhofer@know-center.at

The KNOWSKITE-X Project

Development of science-based electrode materials for reversible chemical-to-power cells (fuel cell, electrolytic cell), enabling efficient energy conversion and storage.

Key Benefits:

- Supports renewable energy integration by storing excess power as carbon-free fuel.
- Focuses on **Mixed Perovskite Oxides** with reduced critical content while maintaining high performance and economic viability.

Objective: uncovering correlations between composition, structure, activity, and performance in Perovskite-based electrode materials using **multi-scale modeling**, **advanced characterization**, and **machine learning**.

Problem Formulation

Modeling Mixed Perovskite Oxides is challenging due to

- Data scarcity** in high-fidelity simulations and experimental measurements.
- Compositional complexity** due to the vast number of elements considered in the composition^[1].
- Computational expenses** as first-principles methods, like Density Functional Theory (DFT), are accurate but costly.

Hybrid modeling with Gaussian Process Regression can help in modeling Mixed Perovskite Oxides by:

- Learning structure-property relationships from **limited data**.
- Predicting key material properties, like DFT energies, **accurately & efficiently**.
- Reducing reliance on expensive simulations or experiments.

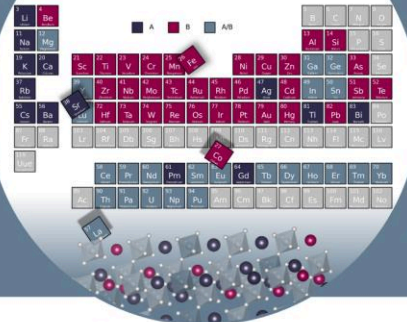
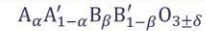
Gaussian Process Regression

- Models target property E as non-linear, continuous function, $E = f(x)$.
- Fits data examples $\{(x^i, E^i)\}_{i=1}^N$ to learn underlying function f .
- Measures similarity through kernel $k(x, x')$ with $k \rightarrow 1$ as $\|x - x'\| \rightarrow 0$.
- Relies on **structure representation** x .

Prediction
(on new data x')

$$E(x') = \sum_{i=1}^N \lambda^i k(x^i, x')$$

Mixed Perovskite Oxides



Structure Representation

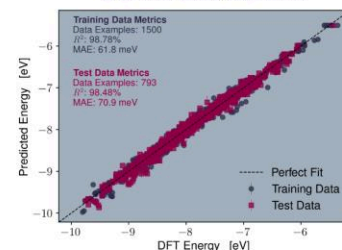
- Uses **descriptors/features** to define space in which similarity is measured.
- Compositional features** by group & period of elements A, A', B, and B'.
- Doping features** by A- and B-site doping levels α and β .

Composite kernel
(to measure similarity)

$$k = \underbrace{k_a}_{\text{doping}} \cdot \underbrace{[k_A + k_A']}_{\text{composition A-site}} + \underbrace{k_b}_{\text{doping}} \cdot \underbrace{[k_B + k_B']}_{\text{composition B-site}}$$

Hybrid Modeling of DFT Energies

... under Data Richness

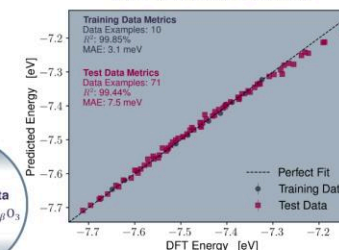


Data mined from the
Materials Project^[2]



Generated Data
 $La_{\alpha} Sr_{1-\alpha} Co_{\beta} Fe_{1-\beta} O_3$

... under Data Scarcity



Funded by the
European Union

- **Challenging data integration & analysis** due to
 - Many partners → **diverse approaches**, many methods → **inconsistent workflows**
 - Much data → **large, complex datasets**, many data formats → **interoperability issues**
 - **Uncertain data relevance** → difficult to decide what should be stored long-term
 - **Complex storage requirements** → need a structured yet flexible collection system
 - **Ease of use vs. complexity** → should be simple despite complex storage rules
- **Potential Solutions** through
 - Shared and adaptive data management platform **based on iRODS**
 - Micromanagement through dedicated **(streamlit) apps**
- **Requirements** needed
 - **Time and Resources**
 - **Steady feedback** from project partners and stakeholders





KNOWSKITE-X

Knowledge-driven fine-tuning of perovskite-based electrode materials for reversible Chemicals-to-Power devices

Thanks for your attention!



knowskite-x.eu

Franz Martin Rohrhofer (Know Center Research GmbH, AUT)



Funded by the
European Union

“Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.”