UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

UNIVERSITÀ DEGLI STUDI DI TRIESTE

XXXVIII CICLO DEL DOTTORATO DI RICERCA IN

APPLIED DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

AREA SCIENCE PARK

Reliable AI in Material Science: A FAIR-by-Design Path from Data to Services

Settore scientifico-disciplinare: **INF/01**

DOTTORANDO
TOMMASO RODANI

COORDINATORE
PROF. FRANCESCO PAULI

SUPERVISORE DI TESI
PROF. STEFANO COZZINI

CO-SUPERVISORE DI TESI
PROF. ALBERTO CAZZANIGA

ANNO ACCADEMICO 2024/2025

Abstract

Materials science faces a dual challenge: transforming legacy archives into Findable, Accessible, Interoperable, and Reusable (FAIR)-compliant datasets through retrospective curation (FAIRification), then establishing prospective workflows that embed FAIR principles from inception (FAIR-by-design). This thesis illustrates a FAIR-by-design path from curated data to deployed Artificial Intelligence (AI) services, addressing challenges in experimental microscopy and spectroscopy.

A FAIRification foundation was established by curating a legacy Scanning Tunneling Microscopy (STM) archive into public datasets with rich metadata and formal provenance. Building on this, the research developed a suite of AI models to enhance experimental data with methods for STM artifact detection and generative restoration, and a dual framework for Near-Edge X-ray Absorption Fine Structure (NEXAFS) signal decomposition based on Deep Learning (DL) and Bayesian approaches.

The research concludes with the deployment of these models as operational open-access services within existing European nanoscience infrastructure. Collectively, the contributions of this thesis provide a reproducible methodology that connects principled data stewardship to the creation of reliable, deployable AI tools for the scientific community.

Contents

List of Figures

List of Tables

List of Abbreviations

AFM Atomic Force Microscopy

AI Artificial Intelligence

AIC Akaike Information Criterion

APE-HE Advanced Photoelectric Effect - High Energy

API Application Programming Interface

AREA Science Park Area di Ricerca Scientifica e Tecnologica di Trieste

AsLS Asymmetric Least Squares (AsLS)

CBIR Content-Based Image Retrieval

CC BY Creative Commons Attribution

CLIP Contrastive Language-Image Pre-Training

CMMD CLIP Maximum Mean Discrepancy

CNN Convolutional Neural Network

CNR-IOM Consiglio Nazionale delle Ricerche - Istituto Officina dei Materiali

CPU Central Processing Unit

DDIM Denoising Diffusion Implicit Models

DDPM Denoising Diffusion Probabilistic Models

DFT Discrete Fourier Transform

DL Deep Learning

DMP Data Management Plan

DOI Digital Object Identifier

ELN Electronic Laboratory Notebook

EOSC European Open Science Cloud

ESFRI European Strategy Forum on Research Infrastructures

FAIR Findable, Accessible, Interoperable, and Reusable

FFN Feed-Forward Network

FFT Fast Fourier Transform

FID Fréchet Inception Distance

FM Flow-Matching

FWHM Full Width at Half Maximum (FWHM)

GAN Generative Adversarial Network

GNU AGPL GNU Affero General Public License

GPU Graphics Processing Unit

GUI Graphical User Interface

HDL Hydrogen Desorption Lithography

HPC High Performance Computing

HTTPS HyperText Transfer Protocol Secure

IaaS Infrastructure as a Service

IDRIN Interoperable Distributed Research Infrastructure for Nanoscience

JWT JSON Web Token

KID Kernel Inception Distance

LCN London Centre for Nanotechnology

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MCMC Markov Chain Monte Carlo

ML Machine Learning

MMD Maximum Mean Discrepancy

ModPoly Modified Polynomial

MSA Multi-head Self-Attention

MSE Mean Squared Error

NEXAFS Near-Edge X-ray Absorption Fine Structure

NFFA Nanoscience Foundries and Fine Analysis

NFFA-DI Nanoscience Foundries and Fine Analysis - Digital Infastructure

NFFA-Europe Nanoscience Foundries and Fine Analysis - Europe Pilot

NN Neural Network

NOMAD Novel Materials Discovery

OCR Optical Character Recognition

OFED Overarching FAIR Ecosystem for Data

ORFEO Open Research Facility for Epigenomics and Other

PaaS Platform as a Service

PNRR National Recovery and Resilience Plan

PROV-DM Provenance Data Model

PSNR Peak Signal to Noise Ratio

PVC Persistent Volume Claims

ResNet Residual Network

RL Reinforcement Learning

SEM Scanning Electron Microscopy

SGD Stochastic Gradient Descent

SNIP Statistics-sensitive Non-linear Iterative Peak-clipping

SNR Signal-to-Noise Ratio

SPM Scanning Probe Microscopy

SR Super-Resolution

SSIM Structural Similarity Index Measure

SSO Single Sign-On

STEM Scanning Transmission Electron Microscopy

STM Scanning Tunneling Microscopy

STRAS Surface Structure and Reactivity at the Atomic Scale

TEM Transmission Electron Microscopy

TEY Total Electron Yield

TIFF Tagged Image File Format

TriDAS Trieste Advanced Data Services

TRL Technology Readiness Levels

UHV Ultra-High Vacuum

URL Uniform Resource Locator

VA Virtual Access

ViT Vision Transformer

VT-STM Variable Temperature Scanning Tunneling Microscope

W3C-PROV World Wide Web Consortium PROV

XAS X-ray Absorption Spectroscopy

XPS X-ray Photoelectron Spectroscopy

Chapter 1

Introduction

This chapter introduces the core concepts of the thesis, outlining the challenges in materials science data management, the role of FAIR principles, and the central research hypothesis that guides this work.

1.1 FAIR-by-Design in AI for Materials Science

The application of AI in materials science is deeply intertwined with the quality and accessibility of data, a challenge addressed by the FAIR-by-design paradigm.

1.1.1 Data management and reproducibility crisis

Materials discovery has undergone profound transformation over the past two decades, evolving from single sample characterization toward high-throughput methodologies generating unprecedented data volumes. Modern synchrotron radiation facilities exemplify this shift: third-generation synchrotrons equipped with high-speed detectors generate terabytes of data daily from individual beamlines [1].
The current data proliferation creates significant challenges for materials researchers. Experimental datasets often lack systematic metadata annotation, relying instead on paper logbooks and informal file naming conventions. Proprietary instrument formats hinder data exchange across research groups. Missing or inconsistent metadata prevents efficient querying and filtering. Quality variability within large archives arising from instrumentation or environmental instabilities complicates automated analysis pipelines.
These organizational challenges contribute to the broader reproducibility crisis affecting materials science. Studies attempting to reproduce published results frequently encounter missing input files, ambiguous parameter specifications, or inaccessible datasets.
The absence of standardized data management protocols limits the scientific community's ability to validate findings, extend previous work, or apply Machine Learning

(ML) methods requiring carefully curated training data.

1.1.2 AI application and FAIR data

AI and ML have emerged as transformative tools for accelerating materials science innovation. However, most AI applications in materials science focus on property prediction tasks, and while valuable, this prediction-centric paradigm neglects the broader experimental workflow where AI could provide substantial benefit [2].
Materials science datasets present unique challenges distinct from natural image or text data dominating mainstream ML. Experimental images contain specialized artifacts that are absent from photographic datasets. Physical constraints provide domain knowledge that generic models cannot exploit. Data scarcity, particularly for pristine reference images or spectra, necessitates physics-informed augmentation strategies.
Addressing these challenges requires adherence to FAIR principles, which provide a framework for improving scientific data management [3]. Findability requires persistent identifiers, such as the Digital Object Identifier (DOI), and comprehensive metadata enrichment. Accessibility ensures data retrieval through standardized protocols and Interoperability mandates vocabularies and formats shared within scientific communities. Reusability demands detailed provenance documentation and explicit licensing.
Current implementation of FAIR principles in materials science remains limited. Many research groups apply FAIR guidelines retroactively during data publication, attempting to reconstruct metadata months or years after experiments have concluded. Alternative FAIR-by-design methodologies embed data management practices from project inception, automatically capturing metadata during experiments and maintaining provenance records throughout the research lifecycle.

1.1.3 Research hypothesis

This dissertation argues how FAIR-by-design approaches to metadata enrichment, automated quality control, data enhancement, and interpretability can integrate AI methodologically across experimental techniques, data modalities, and sample materials, enabling operational services deployed within European and Italian nanoscience infrastructure projects. This approach represents a shift from the prediction-focused paradigm toward a data-centric workflow that spans the full research lifecycle, from acquisition to dissemination.
This thesis adopts a complementary perspective by applying ML to the critical upstream prerequisites of data curation, automated quality control, and image restoration that must be resolved before prediction becomes feasible. While research based on prediction assumes availability of high quality and organized datasets, these assumptions often unmet in real experimental settings, which are instead addressed

first in the AI workflow presented.
This thesis contributes research methods and deployed services to the European and Italian nanoscience infrastructures, supported by the computational infrastructure of the Open Research Facility for Epigenomics and Other (ORFEO) datacenter at Area di Ricerca Scientifica e Tecnologica di Trieste (AREA Science Park). Specifically, it involves experimental laboratories and research data within the Nanoscience Foundries and Fine Analysis - Europe Pilot (NFFA-Europe) and Nanoscience Foundries and Fine Analysis - Digital Infastructure (NFFA-DI) projects and an external collaboration with the London Centre for Nanotechnology (LCN).

1.2 Research Challenges and Research Questions

This thesis addresses three research questions spanning different research data stages, namely curation, analysis, and services. These challenges constrain materials science throughput and limit AI adoption in experimental workflows.

1.2.1 FAIRification

This first domain addresses the foundational challenge of unstructured, 'un-FAIR' legacy data. Large experimental archives accumulated over years of research frequently suffer from inconsistent organization and inadequate metadata. The process of curating raw unorganized collections of data according to the FAIR principles is slow, expensive and is specific to the methodology, experimental setup and management guidelines of the acquisition laboratory. The application of AI or other automation methods to unburden researcher from such activities is desirable but of difficult application for such challenges. This leads to **Research Question 1 (RQ 1):** How can FAIRification activities be performed on unstructured legacy datasets with minimal impact to the laboratory activity?
In the context of this dissertation, this question is contextualized within the activity of the Consiglio Nazionale delle Ricerche - Istituto Officina dei Materiali (CNR-IOM) laboratory of STM, which stored approximately $420,000$ images recorded over two decades, stored in proprietary binary formats with minimal structured metadata beyond instrument parameters. Critical scientific metadata, such as sample composition and experimental settings, exist only in paper logbooks using informal notation. Manual reconstruction for this large image collection proves infeasible, creating barriers to dataset publication and reuse.
The first objective addresses foundational data management: developing automated workflows enabling FAIR-compliant publication of large experimental STM datasets while minimizing manual annotation burden. The approach combines automated metadata extraction from proprietary instrument formats, World Wide Web Consor-

tium PROV (W3C-PROV)[4] provenance tracking documenting complete data lineage, and material category labeling based on ML image retrieval with expert verification.

1.2.2 Enhancing Experimental Data with AI

The second domain focuses on leveraging AI to improve and analyze research data within the laboratory's activities. Researchers must contend with data quality issues, fundamental experimental limitations, and time-consuming analysis bottlenecks. A common thread uniting these problems is the scarcity of pristine, high-quality, and fully labeled data, which further hinders AI integration.

This leads to **Research Question 2 (RQ 2):** How can AI methods be developed and applied to automate quality control, accelerate experimentation, and enhance analysis within specific experimental workflows, especially when constrained by limited high quality training data and domain-specific physical principles?

In this thesis, this broad question is addressed through three specific objectives targeting different challenges in experimental workflows:

Quality Control: STM images contain experimental artifacts arising from tip contamination and scanner instabilities. Multi-tip artifacts, where the tip apex develops multiple atomic protrusions creating duplicated signals, represent a particularly challenging quality control problem. Manual identification requires expert knowledge and becomes impractical for large datasets. Degraded images are usually discarded, creating a severe class imbalance for supervised learning. The **first objective** addressing this part of RQ 2 is to develop Discrete Fourier Transform (DFT) features combined with Vision Transformer (ViT) architectures for multi-tip artifact detection despite severe class imbalance. The approach exploits frequency domain signatures characteristic of spatial duplication patterns.

Experimental Enhancement: STM imaging involves fundamental trade-offs between acquisition speed and image quality. Constant-current mode operation produces high quality topographic images but requires slow scanning to avoid feedback errors. Faster scanning causes degradation through tip blurring, scan-line noise, or sudden tip-change events. Existing ML approaches for restoration suffer from limitations, such as simplified noise models and a need for large training datasets that are unavailable in typical laboratories. The **second objective** addressing this part of RQ 2 is to pursue computational enhancement by developing generative models enabling STM image restoration and Super-Resolution (SR) through physics-informed synthetic data generation. The approach transforms small pristine reference sets into large synthetic training corpora using accurate forward models of degradation mechanisms, enabling data-efficient learning while respecting physical constraints.

Analysis Automation and Interpretability: NEXAFS spectroscopy provides element-

specific electronic structure information critical for materials characterization. However, spectral analysis requires removing background contributions before extracting peak positions and intensities, a task that traditionally involves manual, iterative fitting. This subjectivity is prone to human error and consumes valuable researcher time, creating a bottleneck that prevents automated workflows. The **third objective** addressing this part of RQ2 is to develop complementary ML and Bayesian inference approaches to NEXAFS background removal offering a dual approach that provides fast deterministic output for real-time processing and rigorous posterior distributions for uncertainty quantification.

1.2.3 Data Services

The final domain addresses the challenge of translating these AI methods from research prototypes into reliable, accessible, and FAIR-by-design services for the scientific community. Developing a model is often only the first step; ensuring it is usable, discoverable, and sustainable within existing research infrastructures is a significant challenge in itself.

This leads to **Research Question 3 (RQ 3):** How can AI models be deployed as reliable services that integrate into the existing digital infrastructure and data policies of the research community?

In the context of this dissertation, this is contextualized by the practical need to provide operational tools for the European and Italian nanoscience infrastructures, namely the NFFA-Europe and NFFA-DI projects.

The **final objective** ensures accessibility: deploying developed methods as production-quality web services and Jupyter notebooks within these nanoscience infrastructure projects, maintaining FAIR principles throughout. This objective completes the data-to-service workflow, indicating sustainable pathways for material science research applications.

1.3 Methodology and Scope

This section details the intellectual framework, practical implementation, and boundaries of the research presented in this dissertation.

1.3.1 Methodological Approach

This thesis title *"Reliable AI in Materials Science: A FAIR-by-Design Path from Data to Services"* requires clarification regarding the relationship between *path*, *pipeline*, and *methods*. The thesis presents a set of methods that address distinct challenges rather than a fully orchestrated pipeline of automated tasks. Each method embodies

FAIR-by-design principles during development, but full workflow integration of the experimental data lifecycle represents future work directions.
The *path* in the title refers to the methodological progression from addressing the foundational FAIRification challenge (**RQ 1**), through the FAIR-by-design development of AI methods for quality control, enhancement, and analysis (**RQ 2**), and culminating in the deployment of FAIR-compliant data services (**RQ 3**). This progression illustrates how to build reliable AI services for materials science, not as a single integrated system but rather a reproducible methodology applicable across experimental domains.
This FAIR-by-design approach is adopted as a set of guiding practical principles applied whenever possible throughout the thesis: data and metadata are processed into open formats to ensure interoperability; code and models are published to ensure reproducibility and reuse; datasets are published in open repositories for long-term accessibility; finally, all outputs are shared under permissive licenses to maximize community uptake and impact.

1.3.2 Scope

This thesis aims for methodological generalization, applying its approach across multiple material systems and experimental techniques rather than focusing on a single model system. The scope is therefore defined by the specific, representative contexts chosen to address each Research Question.
The scope for **RQ 1** and the quality control objective of **RQ 2** are focused on a large, legacy STM archive from CNR-IOM. The experimental enhancement objective of **RQ 2** is addressed using a pristine reference STM dataset from the LCN, while the analysis automation objective of **RQ 2** is addressed using NEXAFS spectroscopy data of perovskite oxides from the Elettra synchrotron in Trieste. Finally, the scope for **RQ 3** involves deploying models for both STM and Scanning Electron Microscopy (SEM) data within the NFFA-Europe and NFFA-DI infrastructure projects.

1.4 Thesis structure and outputs

This section outlines the organization of the dissertation and lists the primary research outputs that have resulted from this work.

1.4.1 Outline

The subsequent chapters are structured to directly address the research questions and objectives previously outlined.
Chapter 2 addresses **Research Question 1**, presenting the automated workflow developed for the FAIR-compliant curation of the CNR-IOM STM archive. It details the

methods for metadata extraction, provenance tracking, and semi-automated labeling that provide the data foundation for subsequent chapters.
Chapter 3 tackles the first objective of **Research Question 2**: automated quality control. It presents the development of a Fourier-enhanced ViT for detecting multi-tip artifacts in the curated STM data.
Chapter 4 addresses the second objective of **Research Question 2**: experimental enhancement. It details the development of physics-informed generative models for STM image restoration and SR, designed to be effective even with limited pristine training data.
Chapter 5 addresses the final objective of **Research Question 2**: analysis automation and interpretability. It presents a dual approach to NEXAFS spectroscopy background removal, developing both a Neural Network (NN) for real-time processing and a Bayesian method for rigorous uncertainty quantification.
Chapter 6 answers **Research Question 3**, documenting the deployment of the AI methods from previous chapters as operational services within the NFFA-Europe and NFFA-DI infrastructure. It details the service architecture, FAIR-compliant integration, and the lessons learned from this data-to-service progression.
Chapter 7 synthesizes the contributions of the thesis. It discusses the methodological themes, acknowledges the limitations of the work, and outlines future research directions that build upon these findings.

1.4.2 Research Outputs

The work presented in this thesis has led to the following peer-reviewed publications:

- Rodani, T., Osmenaj, E., Cazzaniga, A., Panighel, M., Cristina, A., & Cozzini, S. (2023). Towards the FAIRification of Scanning Tunneling Microscopy Images. Data Intelligence, 5(1), 27-42.
- Rodani, T., Ansuini, A., & Cazzaniga, A. Enhancing Multi-Tip Artifact Detection in STM Images Using Fourier Transform and Vision Transformers. In ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications.
- Kolev*, N. L., Rodani*, T., Curson, N. J., Stock, T. J. Z., & Cazzaniga, A. (2025). Generative Image Restoration and Super-Resolution using Physics-Informed Synthetic Data for Scanning Tunneling Microscopy. arXiv preprint arXiv:2510.25921. *

*Equal contribution.

Chapter 2

FAIR curation of STM datasets

Data management procedures are fundamental for high-quality research, particularly when dealing with large volumes of scientific data. Metadata and provenance information add value to datasets and enable them to be found, interpreted, reused, and reproduced. Annotating data through specific metadata standards improves their ability to meet the FAIR principles [3].

This chapter presents the activities carried out on a scientific archive of STM images with the objective of organizing them into a more structured dataset from a FAIR perspective [5], [6]. Since the experimental technique has not substantially changed over the last 20 years, the thesis presented in this thesis towards the FAIRification [7] of legacy data is relevant both for the current research activity in STM and for guiding the FAIR-by-design workflow under active development [8].

This chapter details the stages of this FAIRification workflow. Section 2.1 describes the initial dataset extraction from a legacy archive and the development of a semi-automated labeling pipeline that combines domain expertise with ML. Section 2.2 presents the application of the W3C-PROV standard to model the data's provenance, ensuring transparency and reproducibility. Section 2.3 details the publication process for the curated datasets in an open repository, including their formal FAIR assessment. The resulting dataset is central for this thesis, enabling the artifact detection analysis in Chapter 3 and providing the core data for the operational services deployed in Chapter 6. Section 2.4 provides the chapter's conclusions.

2.1 Dataset Curation

The process of curating the legacy STM archive involved several distinct stages, from initial data extraction to a semi-automated, expert-verified labeling workflow.

2.1.1 STM Data Format and Metadata Extraction

STM images in the dataset were recorded by the Surface Structure and Reactivity at the Atomic Scale (STRAS) research group using constant-current and constant-height mode measurements using a Variable Temperature Scanning Tunneling Microscope (VT-STM). Raw data is composed of forward and backward topography scan arrays stored in binary format in files with extensions `.tf0` and `.tb0`, along with a `.par` file containing instrument variables and other information in text format. For some topographic images, related tunneling current images (stored in `.tf1` and `.tb1` files) are also present.

```
;
;               Omicron SPM Control.
;           Parameter file for SPM data.
;

Format                          : 1

Version                         : V2.2
System                          : SCALA

;
; User Information.
;

Date                            : 07.05.10 10:48
User                            : spm22
Comment                         : -

;
; Scanner Description.
;

Name                            : VT1_N.SCA
VGAP Contact                    : TIP

;
; Scanner Area Description.
;

Field X Size in nm              : 30.0000     ;[nm]
Field Y Size in nm              : 30.0000     ;[nm]
Image Size in X                 : 400
Image Size in Y                 : 400
Increment X                     : 0.0750000     ;[nm]
Increment Y                     : 0.0750000     ;[nm]
Scan Angle                      : 0.00000     ;[Degree]
X Offset                        : 182.325     ;[nm]
Y Offset                        : 270.975     ;[nm]
```

Figure 2.1.1: Metadata *.par* file example.

The extraction of instrument metadata from these `.par` files enables structured querying and filtering of the dataset. Each STM image carries 59 instrument metadata fields that provide valuable information about the conditions under which images

were obtained. These metadata are essential for making data findable and accessible according to FAIR principles. Figure 2.1.1 shows the contents of the raw `.par` file from which metadata was extracted.

2.1.2 Manual Labeling of Material Categories

From the full collection of $420,000$ STM images, an initial batch of approximately $110,000$ images recorded in constant-current mode was selected by filtering based on instrument metadata. These images form a reference dataset enriched with the 59 instrument metadata fields described in Section 2.1.1. However, the most crucial information, the structure and composition of the imaged surface, cannot be recorded in an automated way and has been historically registered only in paper logbooks. To obtain this critical metadata, this thesis developed a workflow based on human annotation, ML techniques, and instrument metadata filtering.

The starting point for material category annotation was manual labeling of image groups. Researchers within the STRAS group typically measured samples of the same material within a given day, and considering the typical workflow, it is reasonable to assume samples should be of the same category within limited time periods. Following these assumptions, this thesis tracked the activity of the research group throughout the 20-year period and created a tentative division of the dataset into broad categories.

A total of 188 plots were created, each composed of at most 100 images sampled from individual months of activity. This collection was manually labeled by domain experts, resulting in a division of the dataset into 18 sample material categories, shown in Figure 2.1.2. The monthly activity distribution revealed that approximately 90% of the dataset could be assigned to specific material categories, with the remaining 10% labeled as *mixed* where multiple materials were measured in the same time period. These results shows that while effective, this broad monthly analysis was insufficient for a complete labeling of the dataset, necessitating a more granular and scalable approach.

2.1.3 Deep Learning for Content-Based Image Retrieval

Despite the manual labeling progress described in Section 2.1.2, a complete day-by-day labeling of all 1,470 days in the dataset would be prohibitively time-consuming. To address this scalability challenge, this thesis leveraged Content-Based Image Retrieval (CBIR), a technique that retrieves images based on visual similarity rather than textual metadata. CBIR systems extract feature representations from image content and measure distances between these representations to identify similar images[9]. Modern CBIR approaches employ DL to automatically learn relevant features from large datasets, eliminating the need for manual feature engineering [10].

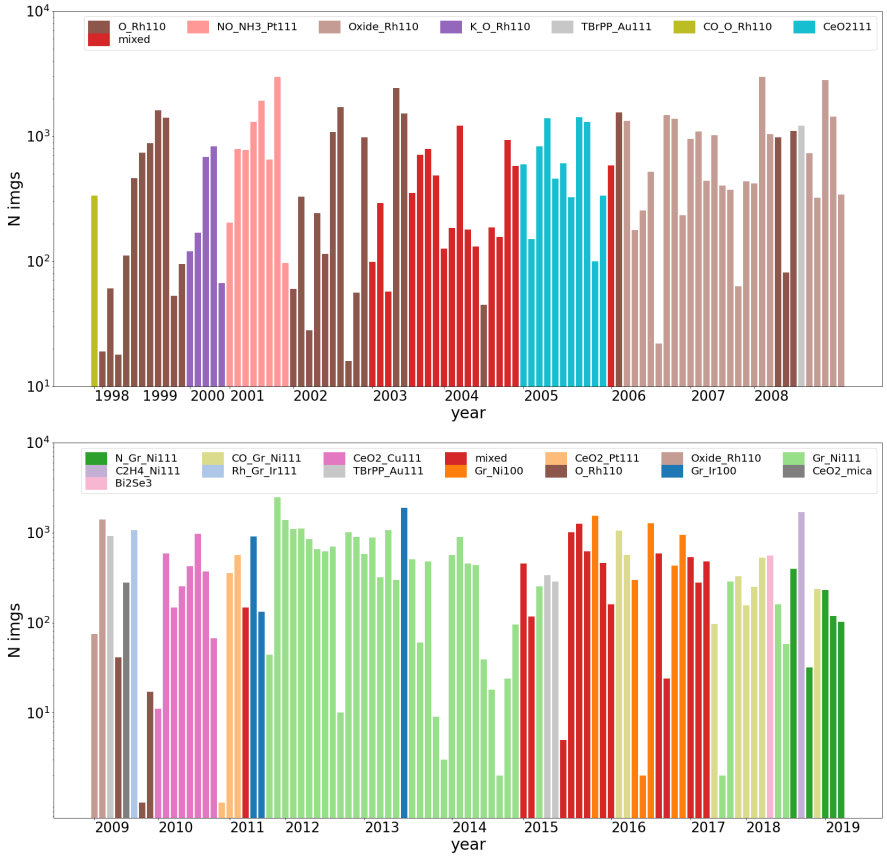

Figure 2.1.2: Monthly activity of STRAS laboratory color coded by sample material category. Months where more than one sample material was recorded are labelled as "mixed".

Given the absence of a sufficiently large set of labeled STM images for training a specialized network, this thesis employed transfer learning with a Convolutional Neural Network (CNN), specifically a Residual Network (ResNet)[11] with 50 layers pretrained on ImageNet [12], a dataset of approximately 1.28 million natural images across 1,000 object categories. CNN apply spatial filters to detect hierarchical patterns, from low-level edges in early layers to high-level semantic structures in deeper layers. ResNet's key innovation is residual learning: rather than learning an arbitrary mapping $\mathcal{H}(x)$, each block learns a residual function $\mathcal{F}(x) = \mathcal{H}(x) - x$ and outputs $y = \mathcal{F}(x) + x$ via identity "shortcut" connections (Figure 2.1.3). By introducing identity shortcut connections that preserve gradient flow, residual architectures enable the optimization of very deep networks and resolve the "degradation" phenomenon wherein deeper plain networks exhibit higher training error than shallower ones [13].

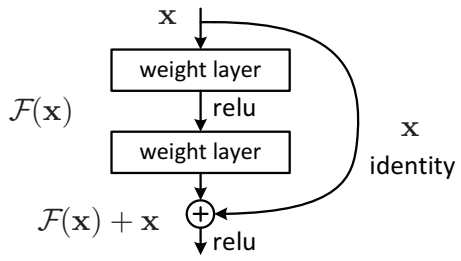

Figure 2.1.3: Residual block: $y = \mathcal{F}(x) + x$. Identity (or projection) shortcuts provide a direct path for information and gradients, mitigating optimization degradation and enabling very deep networks. Figure adapted from He et al [11].

For this thesis, features are extracted from the penultimate layer of ResNet50, yielding a 2,048-dimensional representation encoding image content. Formally, let θ denote the parameters of the NN. This defines a non-linear mapping:

$$f_\theta \colon \mathbb{R}^{224\times224\times3} \longrightarrow \mathbb{R}^{2048} \tag{2.1}$$

which sends each 224×224 RGB image to a compact feature vector suitable for retrieval. The suitability of CNN embeddings for CBIR is well established in the natural image domain, and transfer learning has been shown to be effective in others [14], including SEM [15], [16]. Image similarity is quantified using cosine similarity between feature representations, a standard choice in CBIR for global descriptors [17]. Given two images x_1 and x_2, their cosine similarity is:

$$S_{cos}(x_1, x_2) = \frac{f_\theta(x_1) \cdot f_\theta(x_2)}{\|f_\theta(x_1)\|\|f_\theta(x_2)\|} \tag{2.2}$$

For each image x, the dataset elements where $S_{cos}(x, \_)$ assumes higher values correspond to putative images in the same material category. The semi-automated labeling workflow leveraging this similarity measure is described in Section 2.1.4.

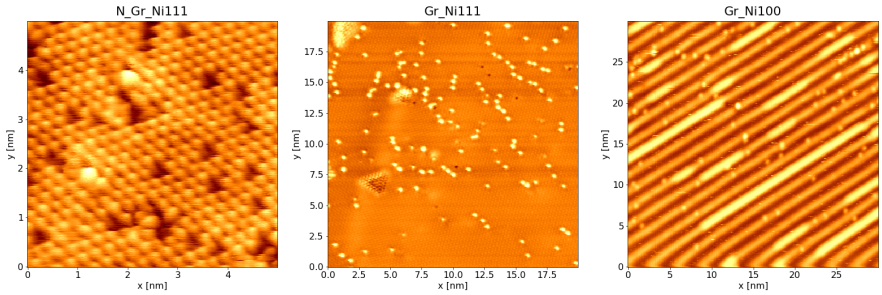

Figure 2.1.4: From left to right: example images of (a) N_Gr_Ni111, (b) Gr_Ni111, and (c) Gr_Ni100 categories from the labeled set of 30 images.

2.1.4 Semi-Automated Labeling Workflow

Based on the representation learning methodology described in Section 2.1.3, a semi-automated labeling workflow was developed consisting of three steps: i) domain experts select a subset of manually labeled representative images for each category; ii) similar images are automatically retrieved using cosine similarity; iii) extracted images are verified by experts. This semi-automated approach drastically reduced the search space, requiring expert verification of a small subset of images (less than 0.2%) to retrieve the final dataset, a task that would have been infeasible with manual inspection alone.

For an initial implementation with three categories, ten representative images were manually selected per category. For each representative image, the 24 nearest images were retrieved using cosine similarity, with constraints to ensure different scan coordinates even if acquisition dates matched. This yielded 720 candidate images, of which 290 unique images were verified as correct by domain experts.

The results highlighted a notable number of duplicates, representing an inefficiency in the workflow that was addressed by modifying the retrieval step to prevent duplicate extractions.

The combination of manual labeling, transfer learning-based image retrieval, and expert verification resulted in a final curated dataset of 7,287 STM images assigned to three material categories of graphene on nickel: Gr_Ni100, Gr_Ni111, and N_Gr_Ni111 as shown in Figure 2.1.4. Here, the numeric suffixes denote the Miller indices (100) and (111) of the underlying single-crystal nickel substrate.

This dataset serves as the basis for subsequent quality control analyses, including the multi-tip artifact detection methodology presented in Chapter 3. Table 2.1.1 details the curation workflow, tracing the derivation of the datasets from the raw legacy archive to the final published collections. These datasets provides an example of how ML techniques can significantly accelerate the FAIRification of legacy scientific data while maintaining quality through expert validation.

Table 2.1.1: Summary of STM datasets in the FAIRification workflow. The table illustrates the hierarchical derivation from the full legacy archive to published datasets.

Dataset Name	Images	Acquisition Technique	Period	Derived From	Description
Raw Dataset	∼420,000	Constant-Current, Constant-Height	1998–2019	VT-STM Measurements	The complete legacy archive acquired by the STRAS group.
Reference Dataset	∼110,000	Constant-Current	1998–2019	Raw Dataset	Filtered subset selected for constant-current mode
Published Reference Dataset [18]	∼71,000	Constant-Current	1998–2010	Reference Dataset	STM images of model surfaces for elementary steps in catalytic reactions, described in 2.3.1.
Curated Labeled Dataset [19]	7,287	Constant-Current	1998–2010	Published Reference Dataset	Dataset labeled with material categories of graphene on nickel, described in 2.3.1.

2.2 Provenance Modelling with W3C-PROV

Provenance is a type of structured metadata that describes the history of data from original sources to final data products. Provenance information tracks all processes applied to data; it is critical for enabling reproducibility and reusability in scientific research. To address these needs, this thesis applies the Provenance Data Model (PROV-DM) [20], a generic data model of the W3C-PROV standard [21].

2.2.1 PROV-DM Core Concepts

The W3C-PROV standard provides a general, high-level framework for describing provenance across diverse domains. The core of PROV is represented by PROV-DM, a generic data model for provenance which defines a common vocabulary allowing interchangeability between systems. The PROV-DM defines three core component types and the relationships between them, depicted in Figure 2.2.1, which provide essential provenance information.

Entities: In PROV, an Entity is defined as *"a physical, digital, conceptual, or other things with some fixed aspects"* [20]. In the STM dataset context, this thesis identified four entities: (i) Raw data, the unorganized collection of $420,000$ STM images acquired using the VT-STM microscope; (ii) Reference dataset, the filtered set of $110,000$ images acquired in constant-current mode; (iii) Structured & FAIR dataset, the $7,287$ images labeled in three material categories; (4) Filtered image, individual images downloadable from the web service STM Metadata Explorer, discussed in detail in Chapter 6.

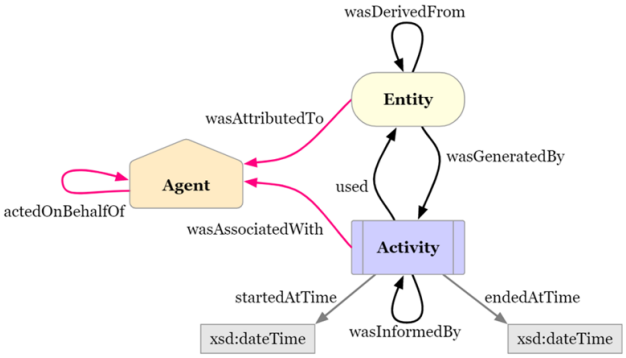

Figure 2.2.1: PROV core structures: Entities, Activities, and Agents, along with relationships between them. Taken from [4].

Activities: An Activity is "*something that occurs over a period of time and acts upon or with entities*" [22]. This thesis mapped four activities in the FAIRification workflow: i) VT-STM measurements, which represent image acquisition from raw data; ii) Image selection & retrieval, the set of actions to obtain the Reference dataset from Raw data; iii) Image labeling process, the pipeline used to enrich the Reference dataset subset with material composition metadata; iv) Metadata selection, the web application workflow to find particular images from the Structured & FAIR dataset.
Agents: In PROV, an Agent [20] can be persons, organizations, software, or other entities that have responsibility for activities or entities. This thesis identified the following agents: i) STRAS research group, which represents the researchers that generated raw data; ii) VT-STM microscope, the instrument used to generate raw data; iii) Data scientist, the agent responsible for the FAIRification process; iv) Research user, the user of the web application interested in the curated dataset; and v) Analysis software, the code used for data processing.

2.2.2 Mapping STM Workflow to PROV

The FAIRification workflow was mapped to PROV relationships following the PROV-DM specification. The workflow begins with VT-STM measurements attributed to the STRAS research group and associated with both the research group and the VT-STM microscope, which acts on behalf of the group. These measurements generated Raw data, which was used during Image selection & retrieval to generate the Reference dataset. The Reference dataset, derived from Raw data, was attributed to both the STRAS research group and the Data scientist. Table 2.2.1 summarizes the mapping between the STM FAIRification workflow elements andW3C-PROV concepts.
The Analysis software acts on behalf of both the Data scientist and Research user. The Image labeling process, associated with the Data scientist and STRAS research group, used the Reference dataset to generate the Structured & FAIR dataset, which

Table 2.2.1: Mapping between the elements of STM case study with W3C PROV concept types and relations

W3C PROV concepts		**STM elements**
PROV types	Entities	• Raw data • Reference dataset • Structured & FAIR dataset • Filtered image
	Activities	• VT-STM measurements • Image selection & retrieval • Image labelling process • Metadata selection
	Agents	• STRAS research group • Data scientist • Research user • VT-STM microscope • Analysis software
PROV relations	Derivation	• Reference dataset **derived from** Raw data • Structured and FAIR dataset **derived from** Reference dataset • Filtered image **derived from** Structured & FAIR dataset
	Usage	• Image selection & retrieval **used** Raw data • Image labelling process **used** Reference dataset • Metadata selection **used** Structured & FAIR dataset
	Generation	• Raw data **was generated by** VT-STM measurements • Reference dataset **was generated by** Image selection & retrieval • Structured & FAIR dataset **was generated by** Image labelling process • Filtered image **was generated by** Metadata selection
	Attribution	• Raw data **was attributed to** STRAS research group • Reference dataset **was attributed to** STRAS research group and Data scientist • Structured & FAIR dataset **was attributed to** STRAS research group and Data scientist • Filtered image **was attributed to** Research user
	Association	• VT-STM measurements **were associated with** STRAS research group and VT-STM microscope • Image selection & retrieval **was associated with** Data scientist • Image labelling process **was associated with** STRAS research group and Data scientist • Metadata selection **was associated with** Research user
	Delegation	• VT-STM microscope **acted on behalf of** STRAS research group • Analysis software **acted on behalf of** Data scientist

was thus derived from the Reference dataset. Finally, Metadata selection associated with the Research user used the Structured & FAIR dataset to generate Filtered images attributed to the Research user. The PROV schematics is shown in Figure 2.2.2.

Part of the terminology used in the provenance workflow has been standardized within the NFFA-Europe community through the MDMC-NEP Glossary of Terms [23]. This standardization effort, coordinated under the NFFA-Europe research infrastruc-

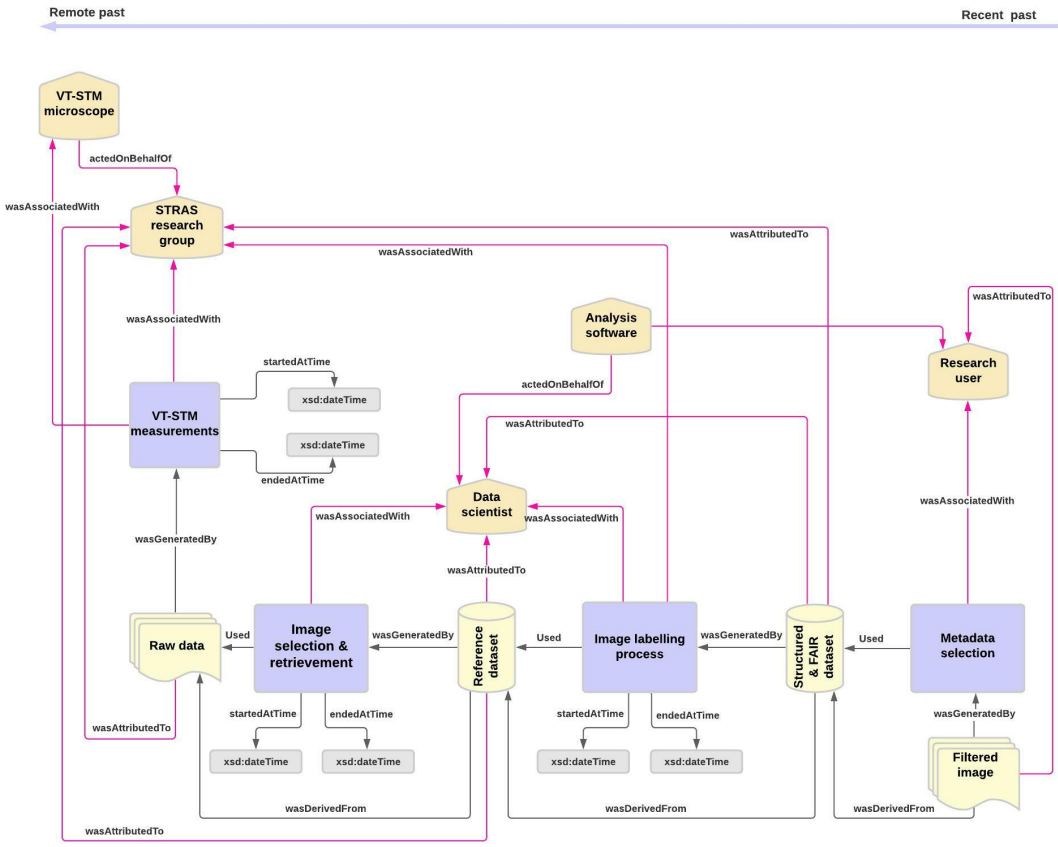

Figure 2.2.2: Graphical representation of the provenance workflow of STM images based on PROV-DM structures. PROV Activities are represented as lilac rectangles, PROV Agents as light orange pentagons and PROV Entities in light yellow ovals. The responsibility properties are depicted in pink. The workflow starts with VT-STM measurements attributed to STRAS research group and is associated with both STRAS research group and VT-STM microscope that acts on behalf of STRAS research group. VT-STM measurements generated Raw data that was used during Image selection & retrieval to generate the Reference dataset. The Reference dataset, which was derived from Raw data, was attributed both to STRAS research group and Data scientist. Analysis software acts on behalf of Data scientist and Research user. The image labelling process, associated with Data scientist and STRAS research group, used the Reference dataset to generate the Structured & FAIR dataset. Therefore, Structured & FAIR dataset derived from Reference dataset. At last, Metadata selection associated with Research user used the Structured & FAIR dataset to generate a Filtered image that was attributed to the Research user.

ture, facilitates interoperability across the broader materials science community and aligns with emerging best practices in research data management.

2.2.3 Implementation and Serialization

The practical implementation of the provenance workflow was conducted using *prov-py*, the PROV Python Library for W3C-PROV Data Model[24]. The provenance document created in Python was exported in PROV-JSON format, providing a compact and

accurate representation that addresses the interoperability and reusability aspects of FAIR principles. The provenance metadata is distributed alongside each image in the dataset, enabling users to trace the complete history from initial acquisition through FAIRification to final publication. This approach aligns with recommendations for nanoscience research infrastructures, where provenance tracking is essential for data quality assurance and reproducibility.

2.3 Datasets publication and FAIR assessment

To maximize the findability, accessibility, and reusability of the STM datasets produced through this FAIRification workflow, this thesis published multiple datasets and associated resources through Zenodo [25], a general-purpose open-access repository developed under the European OpenAIRE program. These publications provide comprehensive access to the curated data, supporting materials, and implementation code.

2.3.1 Datasets publication

Two distinct Zenodo deposits were created to share the outputs of this thesis:

- **STM images of graphene on nickel. DOI: 10.5281/zenodo.7664070**: The primary publication contains 7,287 STM images with complete material category labels (Gr_Ni100, Gr_Ni111, N_Gr_Ni111) [19]. For each image, the deposit includes: STM image data files (`.tf0`, `.tb0` binary format), instrument metadata files (`.par` text format with 59 metadata fields), and W3C-PROV provenance metadata in PROV-JSON format documenting the complete FAIRification workflow. The total dataset size is 3.7 GB.
- **STM images of model surfaces for elementary steps in catalytic reactions. DOI: 10.5281/zenodo.10886977**: A larger reference dataset containing 71,000 STM images recorded in constant-current mode with 59 instrument metadata fields per image, totaling 28.0 GB [18]. This dataset covers measurements from 1998 to 2010 of the batch from which the curated dataset was derived through the semi-automated labeling workflow. The reference dataset enables researchers to explore the full scope of historical STM measurements and potentially extract additional labeled subsets using the published source code.

In addition to the datasets, the complete implementation of the FAIRification pipeline was published as source code on Zenodo [26]. This deposit includes Python scripts for metadata extraction, feature extraction, the CBIR workflow, and the first release of the STM Metadata Explorer web service described in Chapter 6. It contains all dependencies, configuration files, and documentation needed to reproduce the workflow or

adapt it. Together, these publications provide a complete and reproducible record of the STM FAIRification workflow from raw data to curated, labeled datasets. All three Zenodo deposits are released under a Creative Commons Attribution (CC BY) 4.0 International license, which permits sharing and adaptation of the data and code with appropriate attribution. This permissive licensing approach aligns with FAIR principles by removing barriers to reuse while maintaining provenance through attribution requirements.

2.3.2 FAIR Assessment

To verify and assess the level of FAIRness achieved, this thesis employed F-UJI FAIRs-FAIR Data Objects Assessment Service, an open-source tool that supports programmatic FAIR assessment of research data based on a set of core metrics [27].
The FAIR assessment result for the STM dataset was *advanced* level. The dataset achieved full scores on findability (thanks to having a persistent identifier, rich metadata and metadata registered in a searchable resource) and accessibility (having a protocol for metadata retrieval and metadata available even if data is unavailable). The level of interoperability and reusability was rated as *moderate*, indicating areas for future improvement including adoption of domain-specific metadata schemas and enhanced semantic annotations.
Specific areas identified for improvement include: (1) development and adoption of a standardized STM metadata schema; (2) integration with community-accepted ontologies for materials science; (3) enhanced machine-readable licensing information; (4) more detailed usage documentation. These improvements are being addressed through ongoing standardization efforts within the projects described in Chapter 6.

2.4 Conclusions

This chapter detailed the comprehensive FAIRification of a 20-year legacy archive of STM data. The primary objective was to transform an unorganized collection of STM images into a structured and reusable scientific dataset. This was accomplished through a semi-automated pipeline that combined domain-expert annotation with ML CBIR. This approach proved effective in managing the vast scale of legacy data, resulting in a curated dataset of 7,287 labeled images (Section 2.1).
Beyond data curation, this thesis modeled the entire workflow using the W3C-PROV standard. This model (Section 2.2) ensures full transparency and reproducibility by documenting all entities, agents, and activities, from raw data acquisition to final publication. The resulting datasets, source code and provenance records were published on Zenodo under a CC BY 4.0 license. This publication strategy, assessed by the F-UJI service as achieving an *advanced* FAIR level (Section 2.3), confirms the efficacy of

the applied methodology. The impact of published datasets is further evidenced by community adoption, with more than 500 cumulative downloads to date.

The significance of this FAIRification effort is manifested by its direct application in subsequent research. The curated 7,287-image dataset serves as the foundational data for the multi-tip artifact detection analysis presented in Chapter 3.

Furthermore, the curated dataset enabled the development of the STM Metadata Explorer [6], an interactive web service integrated within the Trieste Advanced Data Services (TriDAS) platform. In addition, several other services were built based on the curated datasets and the CBIR methodology described in Section 2.1.3, which were also adapted for other microscopy techniques, such as SEM. The architecture and deployment of these services, which provide a direct link from this chapter's data curation to operational infrastructure, are detailed in Chapter 6.

Future thesis should address the shortcomings noted in this chapter in the FAIR assessments, one of which is being targeted by ongoing efforts: adopting NeXus [28] as a standardized domain-specific format using the NXstm [29] metadata schema [8].

Chapter 3

STM Automated Quality Control

Building upon the curated STM dataset established in Chapter 2, this chapter addresses a common quality control challenge in STM: the automated detection of multi-tip artifacts. While Chapter 2 focused on dataset FAIRification and metadata enrichment, quality assessment of individual images remains crucial for ensuring the scientific validity of subsequent analyses.
This chapter details the methodology and results across three main sections. Section 3.1 defines the multi-tip artifact problem, its physical origins, and the limitations of traditional detection. Section 3.2 details the proposed methodology, including frequency-domain feature engineering, the synthetic data generation pipeline, and the ViT architecture. Section 3.3 presents the experimental validation, providing a comprehensive results analysis, including comparisons to CNNs and ablation studies. Finally, Section 3.4 concludes the chapter and connects this methodology to Chapter 4 and the deployment of this artifact detection model as an operational service in Chapter 6.

3.1 The Multi-Tip Artifact Problem

This section defines the specific quality control challenge addressed in this chapter, detailing the physical origins of multi-tip artifacts and the limitations of traditional methods for identifying them.

3.1.1 Image Artifacts in Scanning Tunneling Microscopy

During STM imaging, the ideal tip terminates with a single atom at the apex, ensuring that only that atom contributes to the measured tunneling current and providing optimal lateral resolution and topographic mapping. When the tip apex changes (e.g., via adsorbates, atom transfer, or mechanical instabilities), artifacts such as broadening, asymmetry and multi-tip can appear[30]. In the specific multi-tip (ghosting) case, the apex ends with two or more atomic-scale appendices that act as independent probes

during scanning. These multiple simultaneous tunneling sites produce a linear superposition of laterally shifted copies of the underlying surface. Visually, repeated structures appear at fixed spatial offsets corresponding to the relative positions of the protrusions [31]. Figure 3.1.1 shows examples of real multi-tip artifacts observed in experimental STM images, illustrating the characteristic duplication patterns at various severity levels.

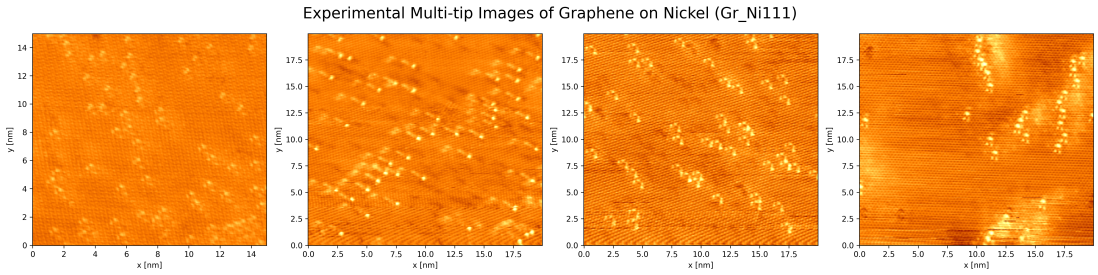

Figure 3.1.1: Examples of experimental multi-tip artifacts in STM images showing characteristic duplication patterns. The artifacts manifest with varying degrees of prominence, from subtle duplications to obvious repeated structures.

The multi-tip artifact presents several challenges for automated analysis, especially compared with other STM artifacts that are detected and mitigated by classical preprocessing and artifacts removal routines [32].
First, multi-tip images exhibit translated duplicates of the true surface which can be mistaken for genuine features, confounding the interpretation of surface composition or crystallographic properties [33].
In addition, the severity varies with the number of active protrusions and their relative heights. When one apex dominates, duplication can be subtle and hard to see directly in real space; when gains are comparable, repeated patterns are more identifiable [34].
Finally, in the Fourier domain, multi-tip duplication introduces additional components at the same lattice frequencies as the primary structure. These often appear as congruent peaks whose phase relationships encode the real-space translation, making it challenging to distinguish ghosting from genuine superstructures [35].

3.1.2 Traditional Detection Approaches and Limitations

Traditional routes link probe morphology to image quality either through analytical simulations of tip–sample interactions [36] or by inferring the probe shape from well characterized surface features [37], [38], [39]. Within the broader Scanning Probe Microscopy (SPM) community, software and methodological frameworks for image quality assessment and artifact characterization are well established, with tools such as Gwyddion[40] widely used for post-acquisition analysis.
While these approaches provide valuable physical insight, they are difficult to scale in practice. Expert visual inspection can identify multi-tip duplication, but becomes

prohibitive for very large datasets such as the one provided in Chapter 2. These workflows delay feedback and preclude adaptive control during scanning, whereas high-fidelity simulations demand detailed models and substantial computation, limiting routine use. Moreover, inverse-imaging strategies rely on specific, known features or calibration structures, which restricts generalization across diverse materials and imaging conditions.

Recent ML efforts detect general adverse tip states, which can include double tip artifacts, with CNNs, line-wise temporal models, and autonomous quality control [41], [42], [43]. These methods, while promising, are often trained on a particular surface and thus tend to learn tip-state cues specific to that surface rather than more general intrinsic features of artifacts. As a result, they may not generalize to new surfaces and are sensitive to class imbalance. Therefore, a significant need remains for a classifier that learns physics-informed features and remains robust across materials for this specific artifact.

3.1.3 Problem Formulation

This thesis formulates multi-tip artifact detection as a binary image classification problem: given an STM image $I \in \mathbb{R}^{H \times W}$ where H and W denote the spatial dimensions, predict whether the image contains multi-tip artifacts (positive class) or represents artifact-free data (negative class). The challenge lies in learning discriminative features that reliably distinguish multi-tip duplications from legitimate surface structures exhibiting periodic or quasi-periodic arrangements.

The dataset imbalance presents a significant challenge: in typical STM operation, artifact-free images vastly outnumber those containing multi-tip artifacts. From the reference dataset, manual annotation identified only 82 images ($\approx 3.9\%$) exhibiting clear multi-tip artifacts among 2080 reviewed images from the curated dataset described in Chapter 2. The low percentage of artifacts is an expected consequence of the operational workflow, where images were saved manually after inspection and multi-tip examples were discarded by expert evaluation. This severe class imbalance requires synthetic data generation and careful evaluation strategies to avoid biased performance estimates.

3.2 Methodology

To address the multi-tip artifact detection problem, this thesis combines classical signal processing and physics-informed synthetic data generation to train DL architectures. The approach leverages the Fourier transform to decompose images into frequency components, exposing structure that is subtle or invisible in real space, and fine-tunes a ViT on a multi-channel representation that fuses spatial and spectral in-

formation.

3.2.1 Frequency Domain Feature Engineering

The Fourier transform provides a powerful mathematical framework for analyzing periodic and quasi-periodic structures in images [44]. This transformation can reveal features that are not noticeable in the spatial domain but become prominent in the frequency domain. For a discrete 2D image $I \in \mathbb{R}^{H \times W}$, the 2D DFT is defined as:

$$F(u,v) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} I(x,y) \exp\left(-2\pi i \left(\frac{ux}{H} + \frac{vy}{W}\right)\right) \tag{3.1}$$

where $u \in \{0, 1, \ldots, H-1\}$ and $v \in \{0, 1, \ldots, W-1\}$ represent spatial frequencies in the x and y directions, respectively. The complex-valued output $F(u,v)$ can be expressed in polar form:

$$F(u,v) = A(u,v) \exp(i\Phi(u,v)) \tag{3.2}$$

where $A(u,v) = |F(u,v)|$ represents the amplitude (or magnitude) spectrum and $\Phi(u,v) = \arg(F(u,v))$ represents the phase spectrum. The amplitude spectrum encodes the strength of frequency components, while the phase spectrum encodes spatial relationships between features.

While the DFT provides the theoretical definition, its direct computation is computationally expensive, scaling quadratically with the input size. The Fast Fourier Transform (FFT) algorithm exploits the symmetry and periodicity of the complex exponential to reduce the computational complexity from $O(N^4)$ to $O(N^2 \log N)$ for an $N \times N$ image. Consequently, within this text, DFT refers to the mathematical transformation, while FFT denotes the specific algorithmic implementation used to generate the frequency-domain features.

Prior to transformation, standard preprocessing procedures are applied: plane-leveling and scan-line alignment to remove global tilt and basic instrumental drift. To exploit information from the frequency domain while preserving spatial features, this thesis constructs a three-channel input representation. The first channel contains the original STM image $I(x,y)$, the second channel contains the amplitude spectrum $A(u,v))$ of the 2D FFT and the third channel contains the phase spectrum $\Phi(u,v) \in [-\pi, \pi]$.

This multi-channel representation enables NNs to learn complementary features across spatial and frequency domains. Figure 3.2.1 illustrates the three-channel representation for a pristine image and its synthetic multi-tip variant, showing how the FFT transformation reveals artifact signatures in the frequency domain.

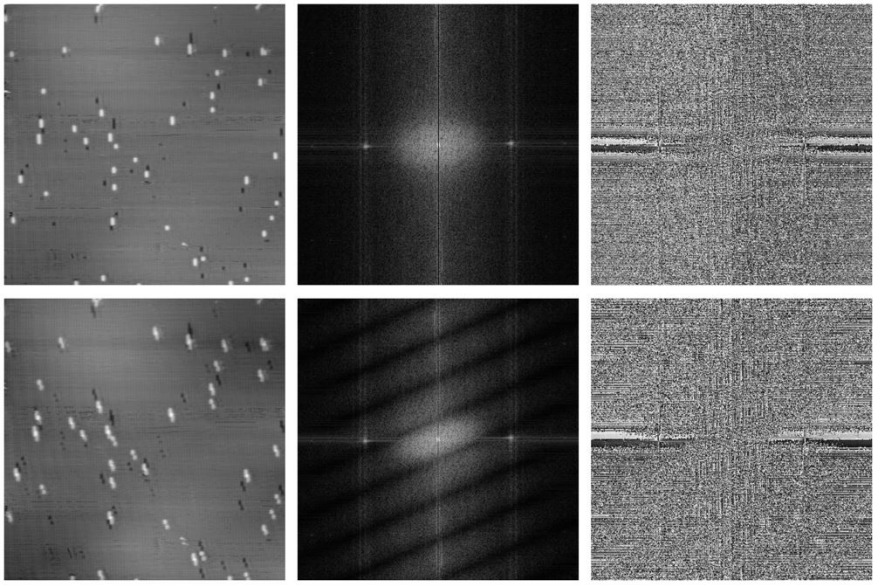

Figure 3.2.1: Three-channel input representation showing grayscale (left), FFT amplitude (center), and FFT phase (right) for a pristine STM image (top row) and the same image after applying synthetic multi-tip artifact (bottom row, synthetic generation described in Section 3.2.2). The amplitude channel reveals characteristic high-frequency patterns induced by the duplication.

3.2.2 Synthetic Dataset Generation

The starting point consists of 2,080 manually annotated STM images of 400x400 pixels from the curated dataset (Chapter 2), including 82 real multi-tip artifacts and 1,998 artifact-free images. All images were acquired using the Omicron VT-STM under varied experimental conditions, including different materials, scan parameters, and environmental settings. Due to the severe class imbalance in real data, this thesis generated a balanced synthetic dataset enabling controlled experimentation. The synthetic data generation pipeline leverages the physics of multi-tip artifact formation while introducing sufficient variability to prevent overfitting.

For an artifact-free image $I(x, y)$ and its multi-tip version with N tips, the relationship can be expressed as:

$$I_{multi}(x, y) = I(x, y) + \sum_{k=1}^{N-1} \alpha_k I(x - \Delta x_k, y - \Delta y_k) \tag{3.3}$$

where $(\Delta x_k, \Delta y_k)$ are translation vectors and $\alpha_k \in [0, 1]$ are attenuation factors accounting for reduced tunneling current from secondary tip protrusions.

Under this translation-superposition model, spatial shifts map to phase ramps in the Fourier domain, yielding congruent sets of peaks at spatial frequencies with consistent phase offsets.

The Fourier amplitude of I_{multi} exhibits interference patterns where the regular spac-

ing $(\Delta x_k, \Delta y_k)$ creates periodic modulation in frequency space, producing distinctive signatures absent in artifact-free images.

Half of the $1,998$ pristine images are transformed into synthetic multi-tip images using the duplication model shown in Equation 3.3, where hyperparameters are sampled from controlled distributions. The number of tips $N \sim \mathcal{U}\{2, 3, \ldots, 12\}$ defines the multi-tip configurations used in this thesis and validated in the ablation studies in Section 3.3. Each component of the translation vectors is independently sampled as $\Delta x_k, \Delta y_k \sim \mathcal{U}\{2, 3, \ldots, 8\}$ pixels, corresponding to atomic-scale tip apex configurations at typical STM magnifications. The intensity factors $\alpha_k \sim \mathcal{U}[0.5, 0.8]$ model reduced current contributions from secondary tips.

Figure 3.2.2 shows examples of synthetic multi-tip artifact generation with varying numbers of tips and translation parameters, illustrating the systematic process that enables controlled experimentation with varying artifact complexity.

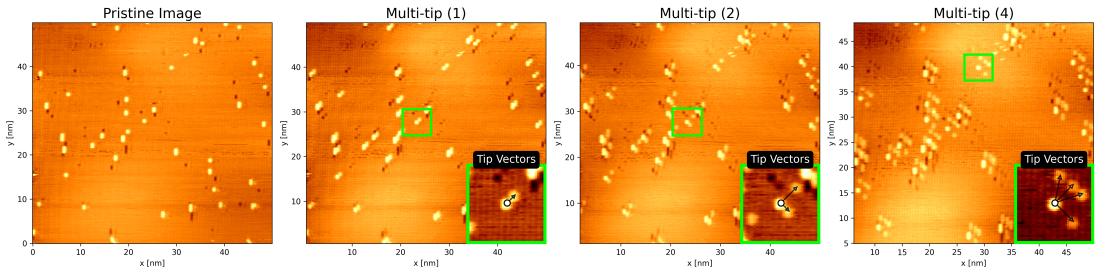

Figure 3.2.2: Examples of synthetic multi-tip artifact generation showing pristine STM images (left) and their synthetic multi-tip variants with different numbers of tips and translation vectors. The systematic generation process enables controlled experimentation with varying artifact complexity.

Images were cropped to 224×224 pixels, the standard input size for ViT models, and augmented with the three-channel representation (spatial, FFT amplitude, FFT phase) as described in Section 3.2.1. Critically, all 82 real multi-tip images were reserved exclusively for the test set, ensuring that evaluation measures generalization to authentic artifacts rather than mere memorization of synthetic patterns.

3.2.3 Vision Transformer Architecture

Traditional CNNs have served as the foundation for computer vision tasks, with architectures like ResNet [11] achieving strong performance through hierarchical feature extraction. However, recent work has indicated that Transformer architectures, originally developed for natural language processing, can achieve state-of-the-art results on image classification tasks when trained on sufficient data [45]. ViT divides an input image into fixed-size patches, linearly embeds each patch, and processes the sequence of patch embeddings through a standard Transformer encoder.

For an input image $I \in \mathbb{R}^{H \times W \times C}$ (where C denotes the number of channels, here

$C = 3$ for the three-channel FFT representation), the image is divided into N patches of size $P \times P$, subject to the constraints $0 < P \leq min(H, W)$:

$$N = \frac{H \times W}{P^2} \tag{3.4}$$

Each patch $I_p \in \mathbb{R}^{H \times W \times C}$, indexed by $p = 1, \ldots, N$, is flattened to $\mathbb{R}^{(P^2 \cdot C)}$ and linearly projected to a D-dimensional embedding where D denotes the constant latent vector size used throughout the Transformer encoder:

$$\mathbf{z}_0^p = \mathbf{E} I_p + \mathbf{e}_{pos}^p \tag{3.5}$$

where $\mathbf{z}_0^p \in \mathbb{R}^D$ denotes the embedding of the p-th patch at layer 0 (i.e., the initial input to the Transformer stack). Here, $\mathbf{E} \in \mathbb{R}^{D \times (P^2 \cdot C)}$ is the patch embedding matrix, flat$(\cdot)$ represents the flattening operation, and $\mathbf{e}_{pos}^p \in \mathbb{R}^D$ is a learnable position embedding encoding the spatial location of the patch. The sequence of patch embeddings is processed through L Transformer encoder layers, each consisting of Multi-head Self-Attention (MSA) and Feed-Forward Network (FFN) blocks. The MSA mechanism computes attention weights between all pairs of patches, enabling the model to capture long-range dependencies. For each attention head h, queries $\mathbf{Q}^h$, keys $\mathbf{K}^h$, and values $\mathbf{V}^h$ are computed from the input embeddings:

$$\text{Attention}^h = \text{softmax}\left(\frac{\mathbf{Q}^h (\mathbf{K}^h)^T}{\sqrt{d_k}}\right) \mathbf{V}^h \tag{3.6}$$

where $(\cdot)^T$ denotes the matrix transpose operation. The scaling factor $\sqrt{d_k}$ is derived from d_k, the dimensionality of the key and query vectors (typically $d_k = D/H_{heads}$, where H_{heads} is the number of attention heads). This scaling prevents the dot products from growing too large in magnitude, which would otherwise push the softmax function into regions with extremely small gradients. The outputs from multiple heads are concatenated and linearly projected. This attention mechanism enables ViT to adaptively weight the importance of different image regions for classification.

Recent theoretical analysis suggests that multi-head self-attention inherently performs low-pass filtering on image signals [46], potentially leading to rank collapse and loss of high-frequency information in deep ViTs. The FFT-based preprocessing counteracts this tendency by explicitly preserving high-frequency components in the amplitude channel. These high-frequency features, crucial for detecting subtle multi-tip duplications, remain accessible to the attention mechanism despite the low-pass filtering characteristic. This thesis fine-tunes pre-trained ViT models, specifically ViT-Base with patch sizes 16 and 32 (denoted ViT-B/16 and ViT-B/32), on the synthetic multi-tip dataset. Pre-training on ImageNet provides strong initialization for general visual feature extraction, which transfer learning adapts to the specific characteristics of STM image artifacts.

For comparison, this thesis also evaluates ResNet architectures (ResNet18 and ResNet50), introduced in Chapter 2, under identical training conditions. The hierarchical feature extraction differs fundamentally from ViT's global attention mechanism. A key distinction relevant to multi-tip detection lies in frequency response characteristics. ResNet architectures exhibit vulnerability to high-frequency noise, as convolutional filters can amplify or suppress specific frequency ranges depending on learned weights. In contrast, ViT's self-attention mechanism reveals greater robustness to high-frequency perturbations, as evidenced by experimental results in Section 3.3. The combination of FFT preprocessing, which highlights high-frequency artifact signatures, and ViT's robust handling of frequency content creates a synergistic effect absent in CNN-based approaches.

3.3 Experimental Validation

This section presents the experimental setup, quantitative results, and ablation studies analyzing the FFT-enhanced ViT approach for multi-tip artifact detection. The evaluation indicates that the proposed method achieves high accuracy on real experimental data while providing insights into the contribution of individual components and the importance of physically informed synthetic data generation.

3.3.1 Classification Results

The dataset was partitioned into 85% for training and 15% for testing, with the training set further divided into 90% for training and 10% for validation. All models were fine-tuned using transfer learning from ImageNet pre-trained weights. Training hyperparameters were selected through preliminary experimentation. The optimizer used was Stochastic Gradient Descent (SGD) with momentum 0.9, and the learning rate was set to 0.01 with no scheduled decay. Cross-entropy loss served as the loss function, and the batch size was 32. Training proceeded for a maximum of 50 epochs with early stopping (patience 10 epochs on validation loss) to prevent overfitting. Significant data augmentation, composed of standard random combinations of transformations including rotation, translation, color jittering, and shearing, was used to prevent overfitting and improve generalization.

Table 3.3.1 presents classification accuracies on the test set for different model architectures with and without FFT-based preprocessing. The results reveal substantial performance improvements when incorporating frequency domain information. All architectures show improved performance with FFT preprocessing, with gains most pronounced for ViT models. ViT-B/32 improves from 78.96% to 97.86%, a gain of 18.9 percentage points, while ViT-B/16 improves from 88.11% to 97.25%, a gain of 9.14 percentage points. ViTs substantially outperform ResNets under both conditions. Even

without FFT, ViT-B/16 achieves 88.11% accuracy, far exceeding ResNet50's 58.84%. With FFT, the gap widens further.

Table 3.3.1: Classification Accuracies with and without FFT Preprocessing

Model	Accuracy (%)	FFT-based
ResNet18	52.43	No
ResNet18	57.07	Yes
ResNet50	58.84	No
ResNet50	65.24	Yes
ViT-B/32	78.96	No
ViT-B/16	88.11	No
ViT-B/16	97.25	Yes
ViT-B/32	**97.86**	**Yes**

Interestingly, ViT-B/32, which uses larger 32×32 pixel patches and thus fewer patches and fewer self-attention operations, slightly outperforms ViT-B/16, which uses 16×16 patches, in the FFT-enhanced setting (97.86% vs. 97.25%). This suggests that for this task, the global receptive field provided by coarser patches captures sufficient information when combined with frequency features.

The substantial improvement of ViT over ResNet can be attributed to ResNet's vulnerability to high-frequency noise combined with ViT's robustness. The multi-head self-attention mechanism in ViT performs implicit low-pass filtering [46], but crucially, the FFT-based preprocessing transforms high-frequency artifact signatures into the amplitude channel where they remain accessible despite attention's low-pass characteristics.

3.3.2 Ablation Studies

To understand which components of the FFT representation contribute most to classification performance and to assess the robustness of the approach to variations in synthetic data generation, this thesis conducted three ablation studies examining individual channel contributions, the effect of tip number range, and the impact of translation vector magnitude.

Table 3.3.2 reports accuracies when using each channel individually. The amplitude channel consistently provides the strongest signal across all architectures. For ViT-B/32, the amplitude channel alone achieves 97.86% accuracy, matching the three-channel performance. This indicates that the amplitude spectrum encodes the critical geometric structure of multi-tip duplications. The phase channel, while providing some improvement, contributes less to classification. This aligns with image processing theory where amplitude spectra capture the magnitude of frequency components, while phase spectra encode spatial positions. Interestingly, the spatial channel alone achieves 88.11% accuracy with ViT-B/16, indicating that ViTs can learn to detect

multi-tip artifacts directly from spatial features when using smaller patches.

Table 3.3.2: Classification Accuracies for individual data channels

Model	Spatial (%)	Amplitude (%)	Phase (%)
ResNet18	52.43	**60.06**	53.04
ResNet50	58.84	**62.50**	51.21
ViT-B/16	88.11	**97.56**	66.15
ViT-B/32	78.96	**97.86**	68.90

Table 3.3.3 examines how the range of tip numbers during synthetic dataset generation affects model performance. This parameterization covers multi-tip configurations where spatial separations range from a few pixels to tens of pixels depending on magnification. However, using up to 12 tips during synthetic generation yields optimal performance (84.76%), with diminishing returns or degradation beyond this point. Adding tips up to 12 helps the models learn complex artifact configurations that while infrequent, improve overall classification task.

Table 3.3.3: Impact of Tip Number Range on ViT-B/16 Performance

Max Tips	Accuracy (%)
4	75.30
8	75.61
12	**84.76**
16	71.34

Table 3.3.4 examines the effect of translation vector magnitude on model performance. Using small translation vectors (2-8 pixels) during training yields optimal performance. Larger translations (10-30 pixels) cause significant performance degradation (85.97%), while very large translations (10-50 pixels) show partial recovery (92.98%). Real multi-tip artifacts typically exhibit small spatial offsets (2-8 pixels at typical STM magnifications), corresponding to atomic-scale tip apex configurations. Training on larger translations introduces a distribution shift in which the model learns to detect patterns at frequencies that do not match real experimental artifacts. The intermediate performance at the 10-50 pixel range may reflect that very large translations produce lower frequency patterns more easily distinguished from artifact-free images, whereas 10-30 pixel translations occupy a difficult intermediate regime. This ablation study underscores the importance of physically informed synthetic data generation, where matching the synthetic data distribution to realistic experimental conditions substantially improves generalization to real artifacts.

Table 3.3.4: Impact of Translation Vector Range on ViT-B/32 FFT Performance

Pixel Range	Accuracy (%)
2-8	**97.86**
10-30	85.97
10-50	92.98

3.3.3 Discussion

This thesis represents a dramatic improvement over manual annotation, which would require significant amount of time for an expert user. The computational efficiency combined with high accuracy makes the method viable for both real-time quality control during data acquisition and retrospective analysis of large archival datasets.
The experimental validation presents several key insights. First, the FFT amplitude channel encodes the critical information for multi-tip detection, validating the frequency-domain hypothesis. Second, ViT outperform CNNs substantially on this task, likely due to their robustness to high-frequency perturbations combined with the explicit preservation of high-frequency features in the FFT representation. Third, physically informed synthetic data generation is crucial, as revealed by the strong dependence of performance on realistic translation vector ranges. Finally, the method achieves deployment-ready performance with inference times compatible with real-time experimental workflows.
While phase contains crucial spatial information, ablation studies (Section 3.3) indicate that amplitude often dominates classification performance, with phase providing complementary cues.

3.4 Conclusions

This chapter addressed the quality-control bottleneck of multi-tip artifact detection in large STM datasets. The thesis formulated the challenge as a binary classification problem, using physics informed data generation to construct a balanced synthetic dataset to train efficiently in low data regimes, such as STM microscopy. The primary contribution is a classification framework combining spatial data with frequency-domain features from the FFT. This three-channel representation, comprising spatial grayscale, FFT amplitude, and FFT phase, was used to fine-tune a ViT-B/32 model.
The method was trained on a physically informed synthetic dataset and validated against a test set containing all 82 available real-world artifact images. The results validates the method's efficacy, achieving 97.86% classification accuracy. Channels analysis confirmed that the FFT amplitude channel provides the most critical features for detection, validating the hypothesis that including frequency-domain information helps models identify the artifacts. This approach showed how ViT models

outperformed ResNet architectures, with ViT-B/32 achieving 32.6 percentage points higher accuracy than ResNet50 when both use FFT preprocessing. Ablation studies confirmed that training data must reflect realistic physical parameters (translation vectors in the 2-8 pixel range) to generalize effectively to real artifacts.
The synthetic data generation methodology and the importance of frequency-domain features are further expanded in Chapter 4, where they are used for post processing tasks in the fabrication regime of STM microscopy techniques.
This research method forms the basis for the STM Artifact Classification Service deployed within the operational NFFA-DI infrastructure described in Chapter 6, illustrating the practical impact of automatic multi-tip artifacts classifation as methodological contribution and, potentially, beyond, such as other artifacts within STM microscopy or other electron microscopy techniques.

Chapter 4

Physics-informed STM image restoration and super resolution

This chapter introduces physics informed DL methods for STM image restoration and SR [47], complementing the automated artifact detection framework developed in Chapter 3. These techniques aim not only to computationally correct degraded images but also to accelerate experimental workflows by enabling reconstruction from faster, lower-resolution scans, thereby extending data utility beyond simple quality control. The chapter is organized into four main sections. Section 4.1 introduces the STM operational challenges and reviews prior computational approaches. Section 4.2 details the methodology, covering the physics-informed synthetic data generation pipeline, the generative model architectures, and the evaluation metrics used. Section 4.3 presents the experimental results, evaluating performance on image restoration, SR ($2\times$ and $4\times$), and computational efficiency. Section 4.4 discusses limitations and future directions and Section 4.5 concludes the chapter.

4.1 Introduction

This section contextualizes the proposed framework by examining the operational trade-offs of STM and reviewing prior computational approaches. It highlights specific limitations in existing methods, particularly regarding data efficiency and artifact modeling, which motivate the physics-informed strategy presented in the final subsection.

4.1.1 STM Operational Challenges: Speed and Tip Degradation

STM, along with related techniques such as atomic-resolution Atomic Force Microscopy (AFM), has been widely adopted for imaging and manipulation at atomic scales since its inception in 1981 and has made possible a range of studies on the nanoscale [48]. Despite significant advancements, STM still struggles with two primary bottlenecks:

the slow rate of image acquisition and the frequent degradation of the tip's quality. Both tip contamination and mechanical deformations can severely degrade imaging quality and attaining stable atomic resolution typically demands extensive manual tip conditioning by highly skilled operators. It also lacks robust automated solutions across varied materials and experimental conditions. Similarly, STM imaging can be orders of magnitude slower compared to methods such as Scanning Transmission Electron Microscopy (STEM). Video frame rate STM exists, but requires specialised hardware [49]. At such high tip speeds, constant height mode is needed, which leads to lower image quality, limiting application to surfaces relatively flat [49]. These limitations significantly hinder the efficiency and scalability of STM, especially in applications requiring high throughput.

The trade-off between imaging speed and quality proves particularly acute in constant-current mode operation—where feedback loops maintain constant tunneling current while scanning, which produces highest quality topographic maps but requires slow scanning to avoid feedback errors. Faster scanning causes image degradation through tip blurring, scan-line noise, or sudden tip-change events. Consequently, researchers face a choice: invest significant time acquiring pristine images or accept degraded data with reduced information content. This fundamental constraint limits STM's utility in time-sensitive applications such as observing dynamic surface processes or enabling high-throughput fabrication workflows [50], [51], [52].

4.1.2 Computational Approaches and Their Limitations

Improving a degraded tip state has been the focus of numerous studies and is a prominent direction within the STM community for improving data collection [42], [43], [53], [54], [55]. One such framework used Reinforcement Learning (RL) to select from six predefined actions, an improvement over random choice [53], [54], but still a limited approximation of the wide range of corrective options available to an STM user. Another study went some way towards embedding human evaluation strategies into their tip evaluation tool [42].

Their system used multiple classification categories to incorporate tip-condition assessment during scanning rather than relying on a complete image. Specifically, they used one network for recent scan lines and a Long Short-Term Memory (LSTM) network to integrate historical data, producing a more context-aware evaluation. However, this richer analysis required a substantially larger data set of 6,167 images of a specific sample material. In contrast, other works proposed methods that reduce the demand for data by semi-supervised labelling [6] or synthetic data generation [56].

Previous computational approaches to image enhancement have focused largely on autoencoder [57] or Generative Adversarial Network (GAN) [58] architectures. While Gaussian noise and scan line misalignments have been introduced [57], critical arti-

facts such as multi-tip distortions and tip blurring were neglected, limiting the scope of application. Other approaches have attempted using GANs for image restoration [58]. Although GANs show excellent results, their training is far from simple and is known to be unstable and difficult to converge [59].
These prior approaches share common limitations highlighting methodological challenges: simplified noise models neglecting complex degradation modes arising from tip-sample physics, training data requirements exceeding typical laboratory dataset sizes, training instability particularly for GAN-based methods, and limited artifact coverage addressing only subsets of degradation modes rather than the full spectrum encountered experimentally.

4.1.3 Physics-Informed Generative Restoration and Super Resolution

This thesis introduces a complementary computational strategy that directly corrects image artifacts. The core contribution is a physics-informed framework that learns to restore images and perform SR using only a minimal set of pristine images. This approach can be used to reduce the frequency of tip-conditioning cycles, shortening experimental runtimes. The reduction in conditioning not only decreases the time spent assessing tip quality but also minimises the need to navigate between different surface areas. This is particularly advantageous in applications where specific regions must be revisited—such as in atomically precise fabrication, since piezoelectric hysteresis, without correction, prevents the same nominal (x, y) coordinates from corresponding to the same physical location after large tip movements.
Importantly, the goal is not to take arbitrary or extremely degraded STM images and reconstruct them at atomic resolution—such an approach would run the risk of inaccuracies due to hallucinated features. Instead, the method is most useful for tasks where the surface is well understood or speed and automation is preferable over perfect accuracy, such as STM applications in Hydrogen Desorption Lithography (HDL), atomic manipulation, or sample navigation.
ML methods are also examined to reduce scan duration by super-resolving low resolution images. In general, for STM to be a viable fabrication method outside the laboratory, it must operate faster—current state-of-the-art HDL requires many hours to produce a single electron transistor [60].
A deep-learning approach is designed to overcome these persistent STM challenges through advanced image restoration and targeted SR. Central to the approach is a realistic, physics-informed data synthesis method that augments high-quality experimental STM data with carefully modelled artifacts such as multi-tip, scan line noise, and tip blunting.
Multiple generative models are systematically evaluated, focusing on Flow-Matching (FM) [61] and Denoising Diffusion Implicit Models (DDIM) [62] due to their ease

of training, rapid inference compared to Denoising Diffusion Probabilistic Models (DDPM) and superior performance [63].
By understanding practical constraints encountered in typical STM laboratories, model performance differences between Central Processing Unit (CPU) and Graphics Processing Unit (GPU) environments are further examined, demonstrating near real-time deployment capability within standard laboratory workflows on commodity hardware.

4.2 Methodology

This section details the technical approach used to generate physics-informed synthetic training data and the specific generative model architectures evaluated for the restoration and SR tasks.

4.2.1 Physics-Informed Synthetic Data Generation

Training NNs requires large datasets, which are uncommon in experimental fields such as STM. To address this, a physics-informed synthetic data generation pipeline was developed that applies realistic instrumental artifacts to high-quality experimental data. All pristine, degraded, and low-resolution experimental images of the Si(001):H surface used in this chapter were provided by the LCN and published on Zenodo [64].
The pipeline builds upon a small experimental dataset of 54 pristine images of the Si(001):H surface. Each images has dimensions of 512×512 pixels, corresponding to 100 nm$\times$100 nm. These were divided into training (36 images), validation (12 images), and test (6 images) subsets. To evaluate the generalisation of the models on real-world data a set of 66 512×512 pixel (100 nm$\times$100 nm) degraded experimental images was curated for the task of image restoration. For the SR task, images were collected at multiple resolutions: 4 at 512×256 pixels (100 nm$\times$100 nm), 3 at 256×128 pixels images (50 nm$\times$50 nm), 4 at 512×128 pixels (100 nm$\times$100 nm), and 3 at 256×64 pixels images (50 nm$\times$50 nm).
The generation process begins by choosing a random image from a training, validation, or test set and normalising it between 0 and 1, after which a sequence of stochastic transformations was applied to simulate common experimental failure modes and increase dataset size.
The synthesis pipeline produced three datasets for the three tasks performed using these initial splits: image restoration, SR of $2\times$, and SR of $4\times$. Each of these has a training set of 20,000 samples, a validation set of 2,000, and a synthetic test set of 2,000. Each sample is a two-channel tensor containing the pristine, high-resolution ground truth and the synthetically degraded image.

In order to train a more robust SR model that works with imperfect data, SR proxies are also synthetically degraded. This means that the models trained for SR have to perform both restoration and SR. Furthermore, to isolate performance on specific degradation types, targeted test sets of 1,000 samples each were generated for multi-tip, scan line misalignment, tip change, and blunt tip artifacts. Equivalent sets were also created for the SR task, along with a low-resolution, degradation-free baseline set. These datasets collectively provide a controlled and diverse foundation for evaluating the effectiveness of augmentation and training strategies, and identifying which degradation types are most difficult to restore.
Each degradation type was parametrised by random variables drawn from prescribed distributions, ensuring a varied and realistic dataset. In general, distribution ranges were selected by quantitative comparison with experimental noise (e.g., the height and shape of scan line noise) or by visual inspection (e.g., the σ used for blunt-tip blurring). In addition, the expression used for the multi-tip artifacts, inspired from chapter 3 was further refined and motivated via a quantum mechanical derivation. The full generation pipeline, in order of application, is as follows:

- **(1) Random Rotation:** Each image was rotated by 0, 90, 180, or 270 degrees with equal probability.
- **(2) Multi-tip artifacts:** Multi-tip effects were simulated by superimposing up to four displaced copies of the clean image, $h(x, y)$. Each copy was modified by a sigmoid function, randomly offset and added back on to the original image

$$f(x,y) = h(x,y) + \sum_{i=1}^{N} K_i\left(\frac{A_i}{1 + e^{\,c_i - d_i h(x-\tilde{x}_i, y-\tilde{y}_i)}}\right) \tag{4.1}$$

  where $N \sim \mathrm{Cat}(\{2,3,4\}; \frac{1}{2}, \frac{3}{10}, \frac{1}{5})$, $c_i \sim \mathcal{U}(5,9)$, $d_i \sim \mathcal{U}(7,10)$, amplitude multiplier $A_i \sim \mathcal{U}(1, 2.5)$, and random offsets of $\tilde{x}_i, \tilde{y}_i \sim \mathcal{U}(1, 11)$. $K_i(\cdot)$ is a kernel applied to the doubled image to simulate a different tip shape from the original and it is selected probabilistically from:

  - Gaussian filter with standard deviation $\sigma \sim \mathcal{U}(1,3)$, with probability 0.3;
  - Median filter with kernel size $k \in \{1, \ldots, 9\}$, with probability 0.4;
  - Random filter with entries drawn from $\mathcal{U}(-0.5, 1)$ and kernel size $k \in \{5, 6\}$, with probability 0.3.

- **(3) Scan line Misalignment:** Horizontal misalignments were introduced with probability 0.3 by shifting individual scan lines by an amount drawn from a Gaussian distribution with $\sigma = 0.8$.
- **(4) Random Crop:** A random crop of 128×128 pixels was extracted.

- **(5) Blunt tip:** To simulate the blurring effect of a blunt tip, a Gaussian filter with $\sigma \sim \mathcal{U}(0.3, 0.6)$ was applied with probability 0.6.
- **(6) Tip change:** Abrupt changes in the tip apex were simulated by blurring the image from a randomly chosen scan line onwards with probability 0.6. This mimics a sudden degradation in instrument resolution. With probability 0.5, a constant offset, Δ, was also added to the scan line where the blurring begins, $\Delta \sim \mathcal{U}(s \times 0.05,\, s \times 0.4), \quad s \sim \mathrm{Unif}(\{-1, 1\})$.
- **(7) Downsample and Upsample:** For SR datasets, downsampling was applied at this stage to avoid interfering with the scan line noise in step 8. Images were downsampled by factors of $4\times$ and $2\times$ in the y-direction only, then upsampled back to their original size using nearest-neighbour interpolation. This one-dimensional downsampling reflects the raster nature of STMs, where the imaging time depends on the number of scan lines and the tip speed. The number of pixels per line does not affect the total length of the scan path. The tip speed should remain constant, since increasing it can degrade image quality by exceeding the feedback loop's response time. By reducing the number of scan lines, the total acquisition time decreases, while the pixel density along each line—and thus the in-line spatial resolution—is preserved.
- **(8) Scan line Noise:** Scan line noise was introduced with probability 0.6. The number of affected lines, m, was sampled from $\mathcal{U}(25, 35)$. Noise segments had lengths drawn from $\mathcal{U}(0, 102)$, and each segment was perturbed by one of three functions:
  - Constant offset from $\mathcal{U}(0, 0.4)$, with probability 0.3;
  - Log-normal function $\mathcal{LN}(\mu, \sigma)$, with $\mu \sim \mathcal{U}(1, 2)$, $\sigma \sim \mathcal{U}(0.5, 1)$, with probability 0.45, representing a sudden tip jump with gradual recovery;
  - Sinusoidal function, with probability 0.25.
- **(9) Normalisation:** Each of the channels is independently normalised to a range of $[-1, 1]$.

Larger images were processed by dividing them into overlapping patches, which were individually restored or upsampled, and then recombined. A small overlap between patches, combined with the application of a squared-cosine window, ensures smoother transitions and minimises visible stitching artifacts.

4.2.2 Generative Models and Baselines

Image restoration aims to learn a mapping $\mathcal{G} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^{H\times W\times C}$ that transforms a degraded image x into an estimate $\hat{y} = \mathcal{G}(x)$ of the ground truth y. Foundational

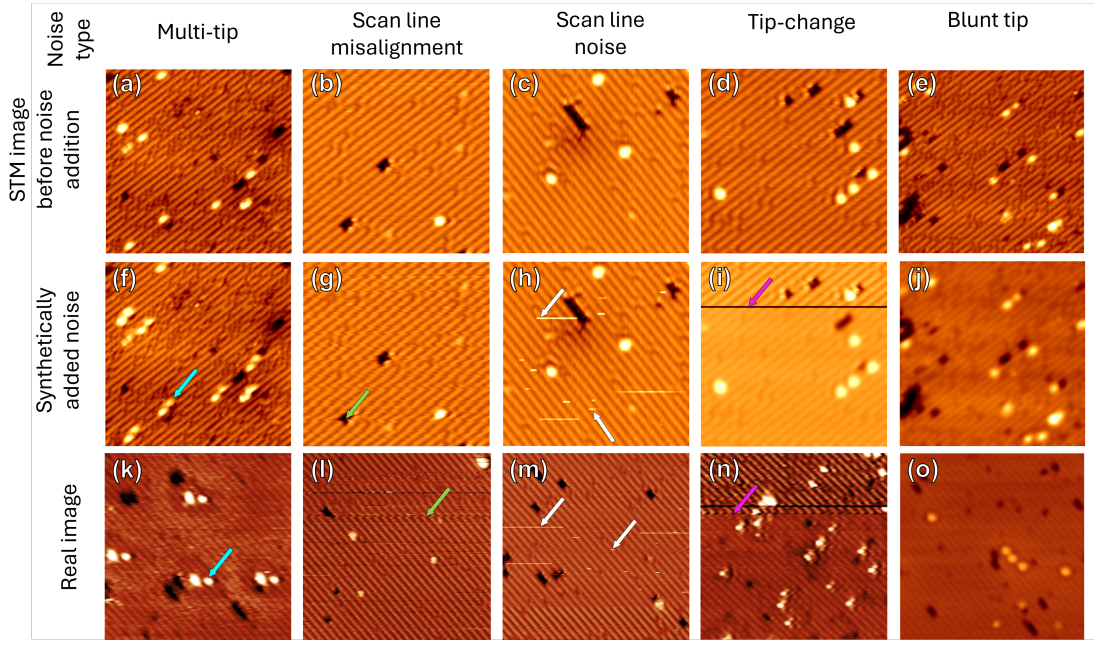

Figure 4.2.1: Comparison of experimental and synthetic STM artifacts. Each image is 25 nm×25 nm (taken at -2 V and between 20 pA and 60 pA). The left column are pristine experimental images, and the middle column are the synthetically degraded pristine images. The right column are degraded experimental images. Each row shows a specific degradation type that we later aim to correct.

architectures for such image-to-image tasks include autoencoders and U-Nets. An autoencoder comprises an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$, where the encoder maps the input to a latent representation $z = \mathcal{E}(x)$, and the decoder reconstructs $\hat{y} = \mathcal{D}(z)$. This compact bottleneck promotes efficient feature learning but often loses fine spatial information. The U-Net architecture addresses this issue through skip connections that link encoder and decoder layers, preserving high-resolution features and improving reconstruction fidelity.

This thesis employs U-Net as the backbone architecture for all generative models, investigating the trade-off between model capacity and computational cost through two configurations: a compact model (3 encoding/decoding blocks, no attention, channel capacities [32, 64, 128], 3.6M parameters) and a larger variant (3 blocks, central self-attention layer, channel capacities [64, 128, 256], 14.7M parameters). The self-attention layer introduced in Chapter 3 enables modeling of long-range dependencies, allowing the network to leverage correlations between distant image regions when reconstructing degraded areas. The autoencoder baseline uses the large U-Net structure but without skip connections, trained with mean absolute error loss ($\mathcal{L}_{MAE}$) to provide a direct comparison against the generative approaches.

Both FM and DDIM share the central idea of defining a forward process in which Gaussian noise is gradually added to a data sample until it becomes indistinguishable

from random noise. A NN is then trained to approximate the reverse process. After training, the model can generate new data starting from random noise and applying the learnt reverse dynamics, progressively reconstructing an image. By conditioning the model with auxiliary information (e.g., a low-resolution input), this mechanism can be adapted to tasks such as denoising, SR, and inpainting [62], [65]. Formally, each $N \times N$ image can be represented as a vector $\vec{x}_0$ in a M-dimensional space, where $M = N \times N$. The distribution of pristine STM images is then defined as $q(\vec{x}_0)$, which the generative models aim to approximate. The details of both approaches are outlined below, including the specific noising procedures, loss functions, and inference methods.

In DDIM, the forward diffusion process gradually adds Gaussian noise to the data in T discrete steps, yielding latent variables $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_T$. Each step is defined by

$$\vec{x}_t = \sqrt{\alpha_t}\vec{x}_0 + \sqrt{1-\alpha_t}\vec{\epsilon}, \quad \vec{\epsilon} \sim \mathcal{N}(\vec{0}, I), \tag{4.2}$$

where α_t is a decreasing sequence from 1 ($t = 0$) to 0 ($t = T$, with $T \approx 1000$ in typical implementations). The NN, $\vec{\epsilon}_\theta$, predicts the added noise, and the training objective is

$$L_{DM} = \mathbb{E}_{t,\vec{x}_0,\vec{\epsilon}}\big[\|\vec{\epsilon} - \vec{\epsilon}_\theta(\vec{x}_t, t)\|\big]. \tag{4.3}$$

The Fourier transform of STM images contains complementary structural information as shown in Chapter 3. Moreover previous studies showed that it improves high frequency detail and edge fidelity in SR and image restoration[66], [67], [68]. An FFT loss therefore complements the pixel-domain training:

$$\begin{aligned} L_{FT,DM} = \tfrac{1}{2}L_{DM} &+ \tfrac{1}{4}\,\mathbb{E}_{t,\vec{x}_0,\vec{\epsilon}}\big[\||F(\vec{x}_0)| - |F(\vec{x}_{t,\theta})|\|\big] \\ &+ \tfrac{1}{4}\,\mathbb{E}_{t,\vec{x}_0,\vec{\epsilon}}\big[\|\arg(F(\vec{x}_0)) - \arg(F(\vec{x}_{t,\theta}))\|\big], \end{aligned} \tag{4.4}$$

where $F(\cdot)$ denotes the Fourier transform, and $\vec{x}_{t,\theta}$ is the reconstruction of $\vec{x}_t$ predicted via $\vec{\epsilon}_\theta$ and Equation 4.2.

Inference reduces to iteratively predicting less noisy states. Crucially, unlike traditional DDPM, DDIM allows for skipping steps: any $\vec{x}_{t-n}$ can be predicted directly, which reduces inference time.

In FM, the forward process is defined by a simpler linear interpolation:

$$\vec{x}_t = \tfrac{t}{T}\vec{x}_0 + \Big(1 - \tfrac{t}{T}\Big)\vec{\epsilon}, \quad \vec{\epsilon} \sim \mathcal{N}(\vec{0}, I). \tag{4.5}$$

Unlike DDIM, which learns a time-dependent denoising function, FM aims to learn a time-dependent velocity field $\vec{v}(\vec{x}_t, t)$ that describes the dynamics between noisy and

clean samples. The training objective is

$$\begin{aligned} L_{FM} &= \mathbb{E}_{t,\vec{x}_0,\vec{\epsilon}}\big[\|\vec{v}(\vec{x}_t,t) - \vec{v}_\theta(\vec{x}_t,t)\|\big] \\ &= \mathbb{E}_{t,\vec{x}_0,\vec{\epsilon}}\big[\|\vec{x}_0 - \vec{\epsilon} - \vec{v}_\theta(\vec{x}_t,t)\|\big]. \end{aligned} \tag{4.6}$$

Where $\vec{v}_\theta(\vec{x}_t,t)$ is the FM network. Inference begins from Gaussian noise, and the learnt velocity field is integrated numerically. The second-order Runge–Kutta midpoint method is used, which provides greater stability and accuracy than Euler integration.

Table 4.2.1: Summary of the model variants tested in this thesis. The models differ in their architectural capacity, determined by channel dimensions and the inclusion of a self-attention layer. Each model is trained with a specific loss function corresponding to its framework: Flow-Matching ($\mathcal{L}_{FM}$), standard DDIM ($\mathcal{L}_{DM}$), DDIM with an auxiliary Fourier-transform loss ($\mathcal{L}_{FT,DM}$), and a mean absolute error loss for the autoencoder baseline ($\mathcal{L}_{MAE}$). The size of the models is given in millions (M) of trainable parameters.

Model	**Parameters (M)**	**Self-attention layer**	**Loss**
Autoencoder	14.7	✓	$\mathcal{L}_{MAE}$
DDIM Small	3.6	×	$\mathcal{L}_{DM}$
DDIM Large	14.7	✓	$\mathcal{L}_{DM}$, $\mathcal{L}_{FT,DM}$
FM Small	3.6	×	$\mathcal{L}_{FM}$
FM Large	14.7	✓	$\mathcal{L}_{FM}$

All models were trained with consistent hyperparameters to ensure fair comparison. Training used the Adam optimizer with a learning rate of 10^{-4}, with a batch size of 32 for 100 epochs, using early stopping to prevent overfitting. The number of inference steps plays a crucial role in both DDIM and FM. Increasing the number of steps typically improves reconstruction quality, but at the cost of higher computational demand and slower image reconstruction. Both frameworks are evaluated across a range of inference steps (e.g. 2, 5, and 10) to characterise the balance between inference speed and reconstruction quality.

4.2.3 Evaluation Metrics

Quantitative evaluation employs two complementary metric categories suited to distinct scenarios. When ground truth data are available, reference-based metrics directly compare reconstructed and target images. For experimental images without ground truth, reference-free perceptual metrics evaluate quality through feature distribution similarity.

Peak Signal to Noise Ratio (PSNR) [69] quantifies the reconstruction quality of an image by measuring the ratio of its maximum possible power to the power of corrupting noise. Expressed in decibels (dB), a higher PSNR value corresponds to a lower level

of reconstruction error.

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{4.7}$$

where MAX_I is the maximum possible pixel value, and MSE is the Mean Squared Error (MSE) between two images.

Structural Similarity Index Measure (SSIM) [70] assesses perceptual image quality by quantifying the degradation of structural information. It models this degradation as a combination of three factors: luminance, contrast, and structure. The resulting index ranges from -1 to 1, where 1 signifies perfect similarity.

$$\text{SSIM}(A, B) = \frac{(2\mu_A\mu_B + c_1)(2\sigma_{AB} + c_2)}{(\mu_A^2 + \mu_B^2 + c_1)(\sigma_A^2 + \sigma_B^2 + c_2)} \tag{4.8}$$

where μ_A and μ_B denote the mean intensities of images A and B, σ_A^2 and σ_B^2 denote their respective variances, σ_{AB} is the covariance between A and B, and c_1 and c_2 are small constants added to stabilize the division when the denominator is close to zero.

For real experimental images without reference data, perceptual metrics assess quality through learned feature representations extracted from pre-trained NNs. The standard models used are InceptionV3 [71], trained on ImageNet [12] for visual features; and Contrastive Language-Image Pre-Training (CLIP) [72], trained on large-scale image–text pairs for joint visual–semantic embeddings. In the following metrics, $A(i, j)$ and $B(i, j)$ represent two $n \times m$ images from sets P and Q, respectively, and α_{net}, β_{net} denote the feature embeddings of $A(i, j)$ and $B(i, j)$ extracted from a pre-trained model net, either CLIP or InceptionV3.

The Kernel Inception Distance (KID) [73] measures the dissimilarity between feature distributions of real and generated images from InceptionV3. Unlike the Fréchet Inception Distance (FID), which assumes Gaussian distributions, KID provides a non-parametric comparison using the squared Maximum Mean Discrepancy (MMD) with a polynomial kernel $k(\alpha_{inc}, \beta_{inc}) = \left(\frac{1}{d}\alpha_{inc}^\top \beta_{inc} + 1\right)^3$, where d is the feature dimension. KID is a more robust and reliable estimator than FID, particularly on smaller datasets.

The squared MMD score is given by

$$\begin{aligned} \text{MMD}^2(P, Q) = {} & E_{\alpha_{inc}, \alpha'_{inc}}[k(\alpha_{inc}, \alpha'_{inc})] \\ & + E_{\beta_{inc}, \beta'_{inc}}[k(\beta_{inc}, \beta'_{inc})] \\ & - 2E_{\alpha_{inc}, \beta_{inc}}[k(\alpha_{inc}, \beta_{inc})] \end{aligned} \tag{4.9}$$

where $E[\cdot]$ is the expectation value.

The CLIP Maximum Mean Discrepancy (CMMD) [74] extends the KID concept using CLIP [72] embeddings. The distance between the feature distributions of real and

generated images is quantified using Eq. 4.9 with a Gaussian kernel $k(\alpha_{clip}, \beta_{clip}) = \exp\left(-\|\alpha_{clip} - \beta_{clip}\|^2/h^2\right)$, with the bandwidth parameter set to $h = 10$. As with KID, a lower CMMD score denotes a higher perceptual similarity.
This evaluation framework employs PSNR and SSIM for quantitative assessment on synthetic data, and KID and CMMD for perceptual evaluation on experimental images, allowing quality improvement to be quantified even in the absence of explicit ground truth.

4.3 Experimental Results

This section presents an evaluation of the physics-informed restoration and SR framework across synthetic and experimental test sets described in Section 4.2. The evaluation is structured into three subsections addressing image restoration performance, SR performance, and computational efficiency. For each task, quantitative metrics (PSNR, SSIM) are presented first to establish reference-based performance on synthetic test data where ground truth is available, followed by perceptual metrics (KID, CMMD) evaluated on real experimental degraded images, and concluding with qualitative assessment of restoration quality and failure modes. For SR, PSNR was omitted due to its known limitations in assessing perceptual quality in this task [75]. To contextualize perceptual metrics, we established empirical bounds: pristine-pristine comparisons set the ideal lower bound, while pristine-degraded (for restoration) or high-res-low-res (for SR) comparisons set the upper bound. Unless specified, results shown represent the best performance across evaluated inference steps for each model configuration. The tables use "FFT" to denote the DDIM variant trained with Fourier loss.

4.3.1 Image Restoration Performance

The quantitative evaluation of model performance on the synthetic test set demonstrates that FM models consistently outperform both DDIM and autoencoder baselines across PSNR and SSIM metrics. Table 4.3.1 summarizes the mean performance across all models.
The FM Large model achieves the highest reconstruction quality with 31.57 dB PSNR and 0.929 SSIM on the synthetic test set. Notably, the difference between FM Large and FM Small is relatively modest.
Analysis of performance across specific degradation types reveals that the models handle artifact classes with varying degrees of success. Table 4.3.2 presents breakdown of PSNR and SSIM on targeted test sets containing isolated degradation types. The models show capability of restoring images affected by a wide range of common STM degradations, although the analysis indicates that multi-tip effects is the most

Table 4.3.1: Image restoration performance on synthetic test set (2000 samples). PSNR measured in dB, SSIM ranges from -1 to 1 (higher is better for both metrics). FM Large achieves the best overall performance with 31.57 dB PSNR and 0.929 SSIM. The number of inference steps significantly impacts reconstruction quality, with diminishing returns beyond 10 steps for DDIM and 5 steps for FM.

Model	**PSNR (dB)**	**SSIM**
Degraded Images	20.00	0.680
Autoencoder	18.82	0.787
DDIM Small	26.59	0.890
DDIM Large	28.86	0.910
FM Small	29.36	0.895
FM Large	**31.57**	**0.929**

Table 4.3.2: Restoration performance breakdown by degradation type on isolated test sets (1000 samples each). Multi-tip artifacts represent the most challenging restoration task across all models, while scan line noise and blunt tip degradation are removed with relative ease. Results averaged across all models.

Degradation Type	**PSNR (dB)**	**SSIM**
Scan line noise	32.38	0.96
Scan line misalignment	30.36	0.94
Blunt tip	32.4	0.95
Tip change	31.04	0.95
Multi-tip	27.18	0.87

challenging to correct, consistent with previous work cited in Chapter 3.
KID and CMMD show a considerable improvement in perceptual quality of the restored images when comparing them to the upper and lower baselines for all models except the autoencoder, as shown in Table 4.3.3.

Table 4.3.3: KID and CMMD scores for restored experimental images (66 degraded experimental test samples). Scores are computed by comparing restored distributions to pristine experimental reference (lower bound) and degraded experimental images (upper bound). Lower scores indicate closer perceptual similarity. All generative models substantially reduce perceptual distance from the degraded baseline toward the pristine reference, though KID and CMMD disagree on relative model rankings.

Model	**KID**	**CMMD**
Synthetically Degraded Images	0.0807	0.446
Autoencoder	0.0830	0.431
DDIM Small	0.0365	0.375
DDIM Large (FFT)	0.0397	0.336
FM Small	0.0331	0.349
FM Large	0.0357	0.350
Pristine Images	0.0194	0.228

All models perform similarly relative to the degraded and pristine baselines, with FM Small achieving the lowest KID score (0.0331) and DDIM Large (FFT) the lowest CMMD score (0.336).
To enable qualitative evaluation with physically meaningful ground truth, a temporal degradation sequence was acquired: the STM tip was first conditioned for stable, high-resolution imaging, after which the same surface area was scanned repeatedly without repositioning. This process induced natural tip degradation over successive scans, yielding a series of images of the same region from pristine to progressively degraded. Figure 4.3.1 provides qualitative evidence from these experimental degraded Si(001):H images.
On the moderately degraded input compromised of scan line noise and misalignment, both models restore scan line noise. FM Large additionally corrects some misalignment while avoiding new artifacts, whereas the Autoencoder produces dark shadowing. On the severely degraded input composed of tip change, gaussian noise and multi-tip, FM Large removes artifacts effectively, though slight bright features elongation occurs; the Autoencoder introduces significant distortions, unnatural global contrast and fails to recover feature shapes.

4.3.2 Super-Resolution Performance

The SR task extends the restoration framework to address acquisition speed limitations by reconstructing high-resolution (512×512 pixels) images from downsampled

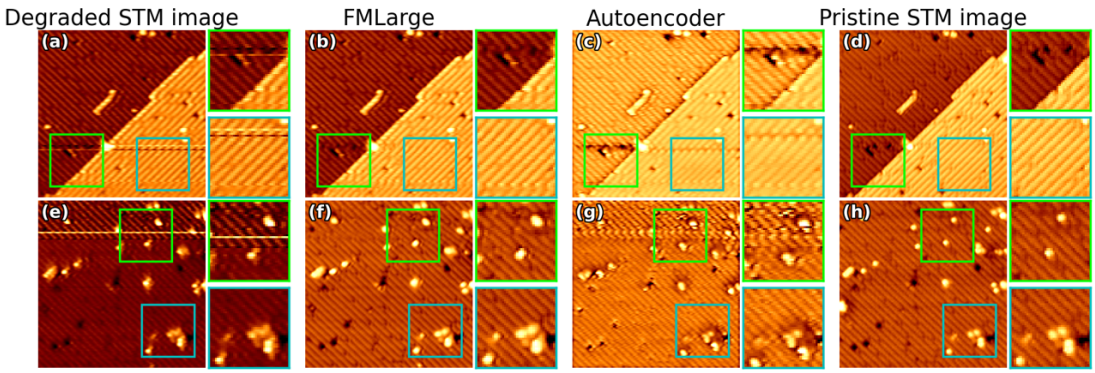

Figure 4.3.1: Qualitative results comparing FM Large with Autoencoder on experimental STM data. Images are 25 nm × 25 nm of Si(001):H taken at -2 V, 20 pA. (a, b) Degraded experimental images exhibiting scan line noise, misalignments, blurriness, tip change, and multi-tip artifacts. (c, d) FM Large restored images with 2 inference steps showing effective noise suppression and multi-tip removal, though with slight feature elongation in severely degraded case (d). (e, f) Autoencoder restored images introducing dark shadowing and unnatural global contrast while failing to recover true feature shapes. (g, h) Pristine experimental ground truth images of the same regions showing natural degradation progression over repeated scans without repositioning.

inputs at 2× and 4× factors. Quantitative evaluation on synthetic test sets demonstrates that generative models effectively perform joint restoration and upsampling as shown in Table 4.3.4.

Table 4.3.4: Super-resolution performance on synthetic test sets (2000 samples per upsampling factor). SSIM evaluated against high-resolution ground truth. Performance degrades gracefully with increasing upsampling factor, with 2× SR achieving near-restoration quality and 4× SR showing modest but acceptable degradation. FM Large demonstrates best overall performance for 2× upsampling.

Model	**SSIM (2×)**	**SSIM (4×)**
Synthetically Degraded Images	0.493	0.396
Autoencoder	0.747	0.672
DDIM Small	0.703	0.661
DDIM Large (FFT)	0.724	0.660
FM Small	0.757	0.719
FM Large	0.778	0.719

Table 4.3.5 presents a breakdown of SR performance across different degradation types applied to low-resolution inputs. The quantitative analysis indicates that most degradation types do not substantially impair SR performance beyond the inherent difficulty of upsampling, with multi-tip artifacts again showing the largest negative impact.

Evaluation on experimental low-resolution degraded images using perceptual metrics is shown in Table 4.3.6.

Table 4.3.5: Super-resolution performance breakdown by degradation type for 2× and 4× upsampling (1000 samples per condition). Multi-tip artifacts represent the dominant adverse condition for SR, whereas scan line noise, misalignment, and blunt tip blur are more readily corrected. Low-resolution baseline (no additional degradation) establishes upper performance bound for each upsampling factor. Results averaged across all models.

Degradation Type	**SSIM (2×)**	**SSIM (4×)**
Low-res Images	0.750	0.723
Scan line noise	0.750	0.703
Blunt tip	0.750	0.713
Misalignment	0.747	0.721
Tip change	0.746	0.723
Multi-tip	0.717	0.672

Table 4.3.6: KID and CMMD scores for super-resolved experimental images. Scores compare super-resolved distributions to high-resolution pristine reference (lower bound) and low-resolution images (upper bound). FM Small achieves lowest KID scores indicating strong perceptual similarity, while DDIM Large (FFT) achieves best CMMD scores for distribution alignment.

Model	**KID (2×)**	**KID (4×)**	**CMMD (2×)**	**CMMD (4×)**
Low-res Images	0.1436	0.1782	0.479	0.601
Autoencoder	0.1021	0.1275	0.407	0.392
DDIM Small	0.1106	0.1359	0.361	0.414
DDIM Large (FFT)	0.1138	0.1020	0.373	0.362
FM Small	0.0977	0.0983	0.401	0.370
FM Large	0.1081	0.0993	0.408	0.362
High-res Images	0.0187	0.0187	0.229	0.229

FM Small yields the lowest KID at 2× and 4× upsampling, whereas DDIM Large achieves the best CMMD at 2× and ties with FM Large at 4×.
Qualitative results in Figure 4.3.2 shows successful reconstruction of fine details like dimer rows and defects for both scale factors.

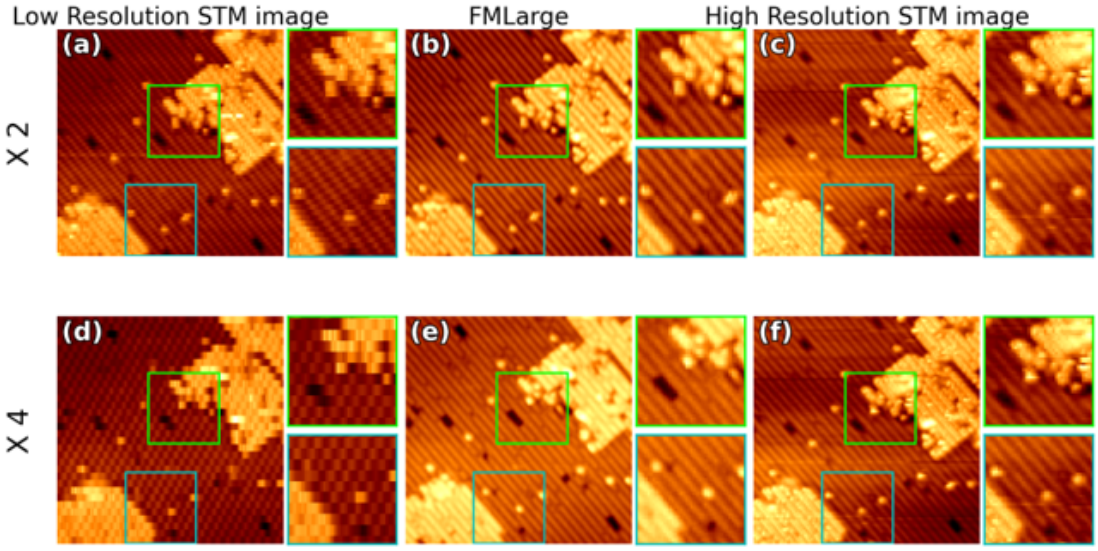

Figure 4.3.2: Qualitative results for 2× and 4× SR on experimental STM images of Si(001):H. Images show same 25 nm × 25 nm area taken at -2 V, 30 pA. (a, d) Low-resolution experimental inputs for 2× and 4× upsampling respectively. (b, e) FM Large super-resolved reconstructions successfully restoring fine atomic details including dimer rows and surface defects. (c, f) High-resolution experimental ground truth images. Insets highlight minor discrepancies in 4× case: small variations in number and location of siloxane defects [76] where effective pixel spacing of low-resolution input approaches characteristic defect size, setting practical upper bound on recoverable detail at high magnification. Despite limitations, the SR pipeline substantially reduces acquisition time while preserving strong structural fidelity.

However, discrepancies in small atomic-scale defects are visible in the 4× case. Degraded low-resolution inputs are reconstructed less accurately than pristine ones, particularly at 4×, where scan-line artifacts can be misinterpreted. The SR framework reliably doubles the resolution of degraded images and accurately reconstructs high-quality inputs well at both scales within physical limits.

4.3.3 Computational Performance and Practical Deployment

A critical consideration for practical laboratory deployment is inference time, which determines whether the framework can be integrated into real-time experimental workflows or requires offline batch processing. Table 4.3.7 summarizes inference times per step for reconstructing a 128×128 pixel patch across model architectures and consumer-grade hardware platforms (AMD Ryzen 5 2600 CPU; NVIDIA RTX 3060 Ti GPU).
DDIM Small offers the fastest CPU inference at 0.10 seconds per step . FM Small

Table 4.3.7: Inference times per step for 128×128 pixel patches on consumer-grade hardware (AMD Ryzen 5 2600 CPU; NVIDIA RTX 3060 Ti GPU). Times shown are average seconds per inference step. DDIM Small achieves fastest CPU inference at 0.10 seconds per step, while FM Small provides good balance of speed (0.21 s/step) and quality. GPU acceleration provides 10-14× speedup across all models.

Model	**CPU (s/step)**	**GPU (s/step)**	**Speedup (×)**
Autoencoder	1.08	0.08	13.50
DDIM Small	0.10	0.01	9.73
DDIM Large	1.13	0.08	13.87
FM Small	0.21	0.02	10.30
FM Large	2.27	0.16	14.08

requires 0.21 seconds per step on the CPU. Leveraging a dedicated GPU provides a significant 10-14× speedup across all models. Crucially, a few steps reconstruction with a small generative model can be faster than a single pass of an Autoencoder to perform a single step. For example, a DDIM Small model with 5 steps completes a reconstruction of a 128×128 pixel image in 0.5 seconds on a CPU (0.10 s/step × 5 steps), compared to 1.08 seconds for the Autoencoder. The proposed SR approach accelerates STM image acquisition rates by 2-4 times, reconstructing sparsely sampled 512×512 pixel images in approximately 11 seconds on standard CPUs. This computational efficiency enables near real-time deployment within standard laboratory workflows on commodity hardware, without requiring specialized GPU devices.

4.4 Discussion

The following subsections analyze the implications of the experimental results, synthesizing the findings and discussing the practical limitations and applications of the proposed methods.

4.4.1 Synthesis and Interpretation of Findings

This thesis demonstrates that physics-informed synthetic data generation combined with state-of-the-art generative models can effectively address STM image degradation and enable SR from minimal pristine experimental data. The central methodological contribution lies in the augmentation pipeline's ability to generate a diverse training set that reproduces the statistics properties of real STM artifacts by modeling their underlying physical origins. This confirms that carefully designed synthetic data can reduce the need for large manually labeled datasets in specialized domains such as STM.

Quantitative results consistently show the superior performance of FM models over DDIM and traditional autoencoders. The strong PSNR and SSIM scores achieved by

FM models, particularly FM Large, indicate a high fidelity in restoring atomic-scale features for scientific interpretation. The relatively modest difference between FM Large and the more compact FM Small suggests that parameter-efficient models can achieve competitive restoration quality, supporting their use in laboratories with limited computational resources.

Evaluation on experimental data using perceptual metrics confirms the effectiveness of the models, with restored image distributions significantly closer to the pristine reference than the degraded inputs or autoencoder outputs. The observed disagreement between KID and CMMD rankings likely stems from the different learned feature spaces and the domain shift between natural and STM images. Although this limits fine-grained model comparison, both metrics consistently indicate substantial perceptual improvement. The preliminary qualitative feedback from expert assessments aligns with these quantitative findings, confirming the physical plausibility of the restored images.

Super-resolution experiments reveal the potential for 2-4$\times$ faster STM acquisition, significantly improving experimental throughput. The ability to reconstruct high-fidelity images from sparse data is particularly valuable for applications prioritizing speed, such as sample navigation or high-throughput fabrication. The computational performance results further demonstrate the practical feasibility of this approach, which supports near real-time integration into workflows without specialized hardware.

4.4.2 Limitations and Application

Despite promising results, several limitations define the scope of current capabilities. Multi-tip artifacts remain the most challenging degradation to fully correct, likely from the inherent physical ambiguity in distinguishing true surface features from duplicated signals.

Super-resolution performance is fundamentally constrained by the input image's effective pixel spacing: features smaller than this limit cannot be reliably recovered, leading to observed discrepancies in atomic-scale defect reconstruction at 4$\times$ magnification. This constraint underscores that SR acts as an accelerator, not a method for generating ground truth at the smallest scales.

On the risk of feature hallucination, qualitative assessments indicate that while models preserve surface characteristics and avoid fabrications in moderately degraded images, reconstructions can become unreliable or even incorrect when input degradation is severe enough to preclude confident human interpretation. Consequently, these methods complement but do not replace proper tip conditioning when quantitative accuracy is required or initial image quality is exceptionally poor.

Therefore, the application context is critical: these models are best suited when speed

and automation are prioritized and the surface is generally well-understood, such as sample navigation, targeting for lithography or manipulation, and initial survey scans. They effectively broaden the range of acceptable raw image quality, increasing the yield of interpretable data from an experimental session.

4.5 Conclusions

This chapter demonstrated the application of physics-informed DL for STM image restoration and SR, addressing key experimental bottlenecks of image degradation and slow acquisition speed. A synthetic data generation pipeline enabled the training of state-of-the-art generative models, starting from a limited set of experimental images.

The results show that these models can enhance STM data quality by restoring common artifacts and achieving 2-4$\times$ acceleration in image acquisition through SR. Their computational efficiency supports near real-time processing on standard laboratory hardware, enabling integration into experimental workflows.

The study also highlights important limitations. Multi-tip artifacts remain difficult to correct, and SR accuracy is constrained by the input image's effective pixel spacing. Accordingly, the methods are most appropriate for applications prioritizing speed and automation, such as sample navigation and lithography, rather than for generating ground-truth data for quantitative analysis. Overall, this thesis establishes a transferable framework for improving scanning probe microscopy workflows and provides validated methods that complement artifact detection (Chapter 3) and can be transitioned to an operational service following the methodology described in Chapter 6.

Chapter 5

NEXAFS Spectroscopy Background removal

This chapter addresses a critical bottleneck in the analysis of NEXAFS spectroscopy data: the removal of complex and variable backgrounds. NEXAFS is a subset of X-ray Absorption Spectroscopy (XAS), a technique that probes the unoccupied electronic states of materials by measuring the absorption of X-rays near the core-level binding energies of specific elements. Focusing on high-throughput *operando* measurements from the Advanced Photoelectric Effect - High Energy (APE-HE) beamline at the Elettra Synchrotron, the thesis compares traditional analysis methods against two complementary computational frameworks. The first framework is a DL approach employing a U-Net architecture trained on physics-informed synthetic data, designed for real-time, automated processing of large datasets. The second framework is a Bayesian Inference via Markov Chain Monte Carlo (MCMC) method that fits explicit physical models to the data, prioritizing rigorous uncertainty quantification and scientific interpretability. Together, these methods balance the speed required for high-throughput screening with the statistical rigor needed in scientific analysis, enabling a more robust and efficient interpretation of spectroscopic data.

5.1 Experimental Context and Background Removal Problem

To understand the challenge of background removal, it is first necessary to describe the experimental setup where the data is generated and the specific nature of the analysis problem.

5.1.1 NEXAFS Spectroscopy at APE-HE Beamline

This thesis focuses on NEXAFS spectroscopy performed at the APE-HE beamline at the Elettra Synchrotron Radiation Facility in Trieste, Italy. The APE-HE beamline covers the photon energy range of 200-1500 eV with variable light polarization enabling

element-specific investigation of K-edges of light elements (C, N, O) and L-edges of 3d transition metals, which are fundamental to the comprehension of the intricate mechanism of material used in applications as catalysis, energy storage, and materials science [77].

The beamline end-station is equipped with a low-noise system for XAS recording in Total Electron Yield (TEY) mode. In TEY mode, the X-ray absorption spectrum is obtained by recording the sample's drain current with a picoammeter while the photon energy is scanned across the absorption edge of the element of interest. This current represents the neutralization response resulting from emitted photoelectrons, Auger electrons, and secondary electrons.

A key capability of the APE-HE beamline is the ability to perform operando NEXAFS measurements at ambient pressure using a specially designed reaction cell [78]. This reactor cell, developed at the CNR-IOM, allows samples to be studied under realistic catalytic conditions with pressure ranges from vacuum 10^{-3} mBar up to 3 Bar, temperature ranges from room temperature to 500°C, and controlled gas environments with multiple gas mixtures [79], [80], [81]. The measurement employs TEY via drain current measurement with a picoammeter, while ultrathin Si_3N_4 membranes serve as X-ray transparent windows that seal the atmospheric sample environment from the beamline Ultra-High Vacuum (UHV).

The reaction cell design enables fast continuous scanning acquisition, allowing full NEXAFS spectra to be recorded in as little as 10 seconds. This rapid acquisition is critical for studying reaction kinetics and dynamic chemical processes in $operando$ conditions [80]. The cell also allows the application of bias voltage between the sample and the Si_3N_4 membrane to increase electron collection and to exploit Townsend avalanche effects in the gas for signal amplification.

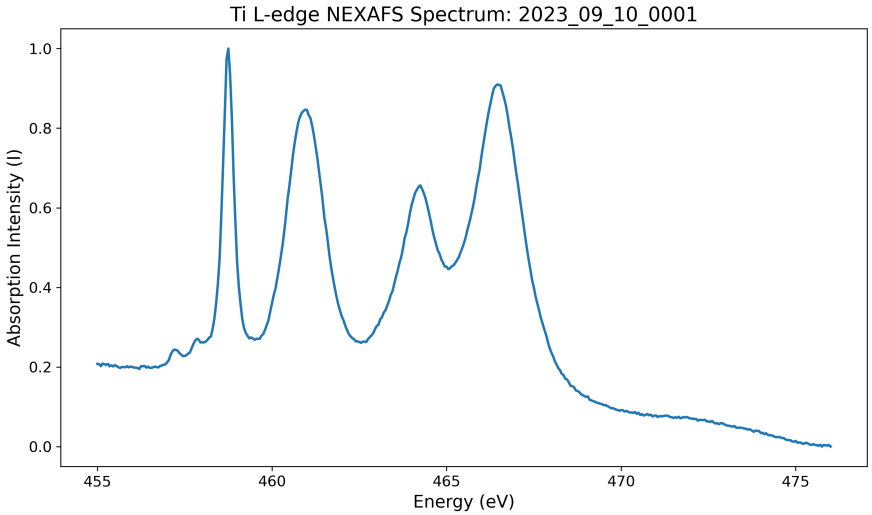

Figure 5.1.1: Ti $L_{2,3}$ edges of lanthanum-doped strontium titanate perovskite ($La_{1-x}Sr_xTiO_3$) collected at the APE-HE beamline.

The primary experimental validation in this chapter focuses on the $Ti\ L_{2,3}$ edges of lanthanum-doped strontium titanate perovskite ($La_{1-x}Sr_xTiO_3$) as shown in Figure 5.1.1. This spectrum was collected at the APE-HE beamline in vacuum at room temperature in TEY mode, representing a characteristic example of the transition metal L-edge spectra commonly measured at the facility. It serves as a reference case for developing and validating the methods presented, which will subsequently be applied to more complex datasets acquired under challenging environmental conditions.

5.1.2 NEXAFS Analysis Workflow and Background Removal problem

Raw NEXAFS spectra from the APE-HE beamline require multi-step preprocessing before quantitative analysis can be performed. The standard workflow begins with energy calibration to correct for monochromator drifts using reference standards, followed by normalization to account for variations in incident beam intensity. At ambient pressure, normalization is affected by absorption from the Si_3N_4 membrane and gas layer, making it more challenging than in traditional UHV measurements. After normalization, background removal separates the slowly-varying or structured baseline from the absorption features of interest, enabling subsequent peak identification and fitting for extraction of quantitative parameters such as edge positions, peak energies, intensities, and widths.

Background removal is a bottleneck in this pipeline, particularly for high-throughput operando experiments where hundreds of spectra may be acquired during a single reaction study. At ambient pressure, the experimental setup leads to complex nonlinear background shapes that are more structured and variable than traditional UHV spectra, increasing the difficulty of automated background removal.

To address this analysis bottleneck, several software packages have been developed. While a mature ecosystem exists for hard XAS with tools like ATHENA [82] and SIXPACK [83], fewer have been specifically designed for the nuances of soft X-ray NEXAFS data, with notable examples being QANT [84] and THORONDOR [85]. These programs provide a suite of functions, often through a Graphical User Interface (GUI), allowing researchers to apply various background subtraction models such as polynomial curves, splines, and Asymmetric Least Squares (AsLS) (AsLS). However, these tools often require manual parameter tuning for each spectrum, hindering true high-throughput automation. Furthermore, their uncertainty estimates can be unreliable. For a more robust analysis, the THORONDOR paper itself recommends more advanced techniques like MCMC sampling as a necessary additional step.

5.1.3 Problem Formulation

From a data analysis perspective, a NEXAFS spectrum is a one-dimensional signal that contains 200-2000 data points consisting of photon energy (E) on the X axis and

normalized intensity on the Y axis.
Background removal is a signal decomposition problem. The measured NEXAFS intensity $I_{\text{meas}}(E)$ is modeled as:

$$I_{\text{meas}}(E) = I_{\text{signal}}(E) + I_{\text{background}}(E) + \eta(E) \tag{5.1}$$

where $I_{\text{signal}}(E)$ represents the desired spectroscopic features, $I_{\text{background}}(E)$ is the background signal and $\eta(E)$ is random measurement noise. The objective is to estimate $I_{\text{signal}}(E)$ only given $I_{\text{meas}}(E)$. This is an ill-posed inverse problem requiring regularization or incorporation of prior physical knowledge to achieve a stable, unique solution.
The key challenges include i) signal-background overlap where spectral features and background are often degenerate in frequency space, preventing simple filtering approaches; ii) diverse morphologies where different sample systems, gas environments, and measurement conditions produce widely varying background shapes; iii) lack of ground truth where the true signal and background are never measured independently, making supervised learning difficult. From a ML perspective, this can be framed as a regression problem where a model learns to predict the continuous background function. The primary obstacle is the scarcity of labeled training data with known ground-truth background separations.
Practical NEXAFS background removal for operando experiments at APE-HE must meet several critical criteria. It should run automatically to support high-throughput operation, and it must process spectra in seconds or less to keep pace with acquisition rates. At the same time, the method needs to preserve peak positions, intensities, and edge features with high accuracy, while delivering reproducible, deterministic outputs for any given input. For critical analyses, it should also provide rigorous uncertainty quantification, supplying reliable error estimates for all extracted parameters.

5.2 Deep Learning for High-Throughput Analysis

This section details the first computational framework, a DL model designed to meet the speed and automation requirements of high-throughput operando experiments.

5.2.1 Motivation and Approach

Traditional background removal methods struggle to meet the throughput demands of modern synchrotron facilities, which can generate hundreds of spectra per hour. Classical automated algorithms often require manual parameter tuning for each spectra, creating significant bottlenecks in the data analysis pipeline. This thesis addresses these challenges using supervised DL. Since experimental spectra lack ground-truth labels for the background, a large scale synthetic dataset was generated.

The synthesis follows standard near-edge practice but remains deliberately general. Peak profiles are modeled as Gaussian, Lorentzian, or Voigt [86], consistent with common near-edge analyses [86], [87], [88], [89]. For example, Voigt broadening is routinely applied in L-edge simulations [90] and Gaussian pre-edge features are treated in Mn L-edge studies [91]. The edge is modeled with an arctangent step, consistent with routine preprocessing and normalization in soft-X-ray workflows [82]. The background is modeled separately from the features, using either a smooth low curvature function (e.g., broad Gaussians/exponentials)[92] or an empirical baseline taken from experimental spectra to capture realistic geometry and detector dependent slopes [93], [94], [95]. Finally, Gaussian noise was added with randomized variance to span realistic Signal-to-Noise Ratio (SNR)s to approximate observed detector noise. Further details on the synthetic data generation are presented in section 5.2.2.
For this task, we selected a 1D U-Net architecture, a model originally developed for biomedical image segmentation [96] whose symmetrical encoder-decoder structure with skip connections has proven highly effective for signal-to-signal translation tasks. U-Net models have been adapted successfully to 1D waveform processing in other domains such as audio source separation [97]. In the spectroscopy domain, ResNet and U-Net hybrids have achieved automated preprocessing and denoising for Raman spectra [98], [99], [100]. One such work employed a pipeline similar to the one outlined here, generating synthetic Raman spectra to train a U-Net model for background separation [101]. As shown in the previous chapter, at inference, a trained U-Net model can process images in real-time on commodity hardware without requiring any manual parameter tuning, making it ideal for this use case. The specific model architecture, training procedure and results are described in section 5.2.3.

5.2.2 Synthetic Dataset Generation

A robust and diverse synthetic dataset is crucial for training a generalizable model. The data generation pipeline models NEXAFS spectra as a linear superposition of physically motivated functional components.
The spectral components include absorption peaks modeled using Voigt profiles, which are convolutions of Gaussian and Lorentzian lineshapes representing combinations of instrumental/inhomogeneous broadening and lifetime broadening effects. The Voigt profile is computed using the Faddeeva function $w(z)$:

$$\text{Voigt}(E; E_0, \sigma_G, \gamma_L, A) = A \cdot \frac{\text{Re}[w(z)]}{\sigma\sqrt{2\pi}} \tag{5.2}$$

where

$$z = \frac{(E - E_0) + i\gamma_L}{\sigma\sqrt{2}}, \tag{5.3}$$

E is the photon energy, E_0 is the peak position, σ_G characterizes the Gaussian width

(related to Full Width at Half Maximum (FWHM) (FWHM) by $\sigma_G = \text{FWHM}_G/2.355$), γ_L is the Lorentzian half-width at half-maximum, and A is the peak amplitude. Absorption edges representing step-like absorption onsets are modeled with arctangent functions:

$$\text{Edge}(E; E_0, \Delta) = \frac{\Delta}{\pi} \arctan\left(\frac{E - E_0}{w}\right), \tag{5.4}$$

where E_0 is the edge position, Δ is the edge jump magnitude, and w controls the edge width.

To generate a variety of smooth background shapes, parameter ranges were chosen empirically. Gaussian backgrounds are generated as broad Gaussian functions. To ensure they form smooth underlying curves and slopes rather than distinct peaks, their centers are randomized to the first half of the spectrum and their widths are set to be significantly larger, between 1.5× and 3× of the spectrum length. Exponential backgrounds follow the form $y = \exp(-x/a)$ where the decay constant a is randomized between 15% and 100% of the spectrum length, as a value below 15% results in an exponential decay that is too steep. These specific parameter ranges were chosen empirically to ensure the generation of smooth, realistic background curves. Real backgrounds are sampled from a curated library of 38 experimental baseline shapes extracted from previous APE-HE measurements. These profiles were isolated by subtracting vacuum reference spectra from raw experimental acquisitions, covering a diverse set of Ti $L_{2,3}$, La, and O K-edge absorption spectra. Additive white Gaussian noise is added to the composite signal, with the signal-to-noise ratio randomized between 10 and 30 dB to simulate realistic experimental conditions.

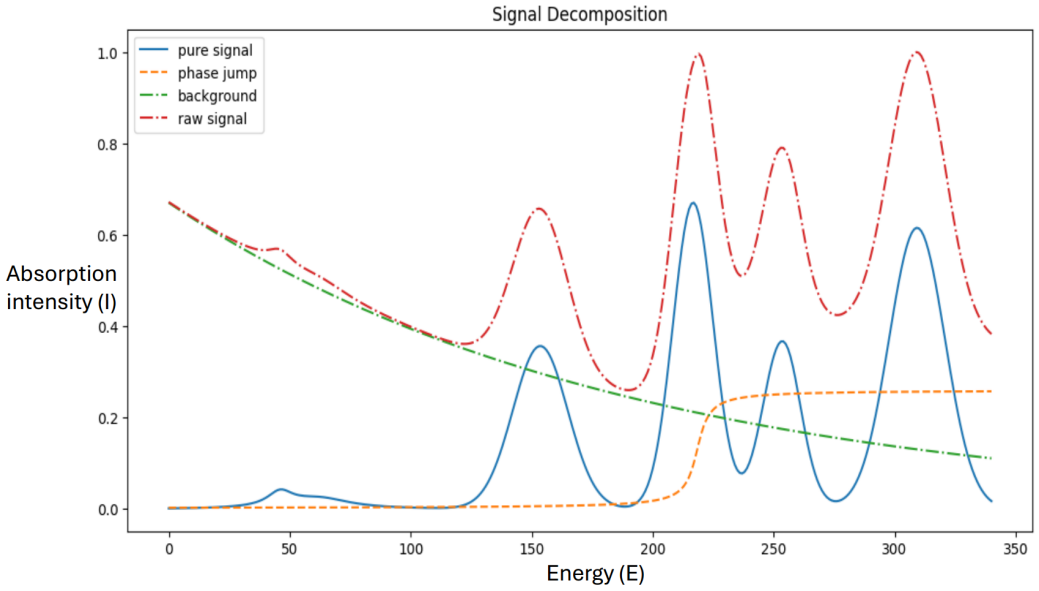

Figure 5.2.1: Synthetic spectra example

Each synthetic spectrum is constructed by first generating a pure signal containing

between 4 and 20 peaks. Peak positions are uniformly distributed across the energy range with a minimum separation of 5% of the spectrum length to ensure distinguishability. Peak heights are randomized between 0.1 and 1.0 (normalized units), Gaussian widths between 1% and 20% of the spectrum length, and Lorentzian widths between 0.5% and 50% of the Gaussian width to produce realistic peak asymmetries. An edge jump with magnitude between 0.1 and 1.0 is added at a position within the first 50% of the energy range. Although the edge position is physically constrained to the first peak, a broader range was used during training to improve model generalization across diverse experimental conditions and ensure robustness to variations in edge placement. Following the same principle, a background is uniformly sampled from either synthetic ones generated from low curvature functions (Gaussian and exponential) or empirically acquired real experimental ones. After adding noise, the entire spectrum is normalized to the range. This process yields paired examples where the input is the noisy composite spectrum and the target is the clean background signal. Figure 5.2.1 show an example of synthetic spectra.
The full dataset consists of 100,000 synthetic spectra, each with 1000 energy points spanning a normalized energy range from 0 to 1. This large dataset size, combined with the diversity of spectral morphologies, ensures that the model encounters a wide range of background shapes and signal characteristics during training.

5.2.3 U-Net Implementation and Performance

The U-Net architecture employed follows the same encoder-decoder backbone family as Chapter 4, adapted for 1D spectroscopic data. The model consists of a symmetric structure with four encoding and four decoding levels (with channel capacities of 64, 128, 256, and 512) for a total of 10,824,129 trainable parameters.
Training employed MSE loss with Adam optimizer (learning rate 10^{-4}, batch size 32) for 50 epochs with early stopping (patience 5 epochs).
The trained U-Net was benchmarked against classical background removal algorithms from the pybaselines library [102] (version 1.1.0): Modified Polynomial (ModPoly)M, AsLS, and Statistics-sensitive Non-linear Iterative Peak-clipping (SNIP). The classical methods were configured with the following parameters: ModPoly with polynomial order 3, AsLS with smoothing parameter $\lambda = 10^7$ and asymmetry parameter p = 0.02, and SNIP with maximum half-window 40, decreasing mode enabled, and smoothing half-window 3.
The evaluation was conducted on a test set of 54 experimental NEXAFS spectra of Titanium measured at the APE-HE beamline. Performance was quantified using MSE, Mean Absolute Error (MAE) and SNR as metrics. Ground-truth backgrounds were constructed by subtracting a vacuum reference from each operando spectrum. Figure 5.2.2 shows one of such backgrounds and the performance of U-Net against classi-

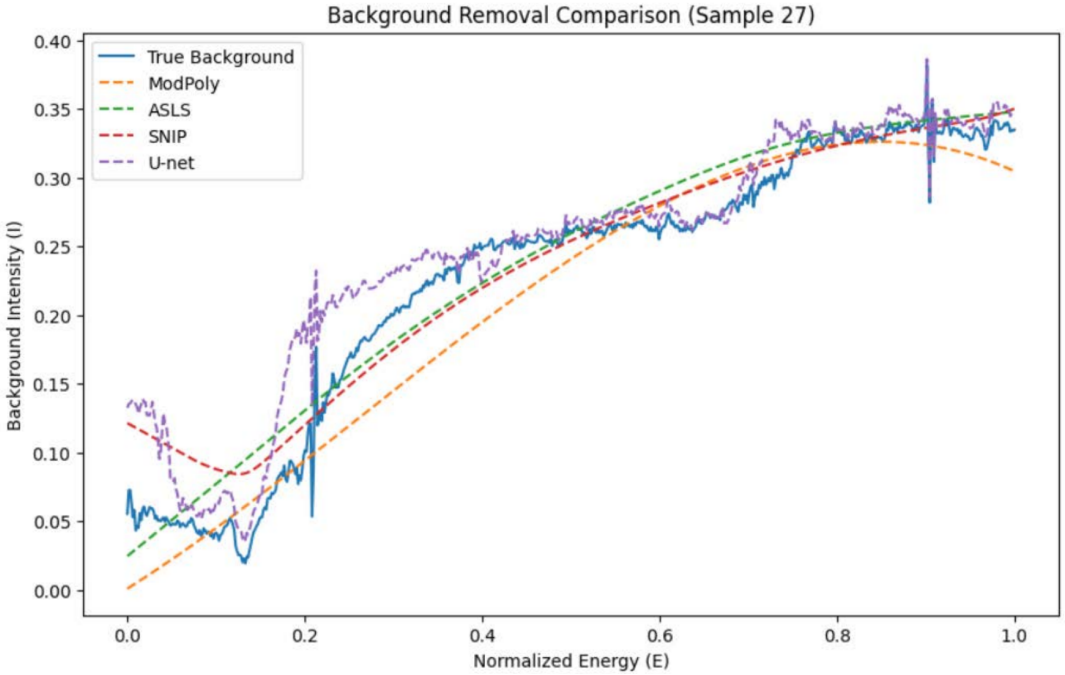

Figure 5.2.2: Background removal of a $Ti\ L_{2,3}$ experimental spectra of U-Net against classical methods.

cal methods. Specifically, given that the measurement were taken in the same session, a vacuum reference on the identical sample was acquired immediately before the operando run using the same beamline settings, detection mode, and energy grid. This "difference-to-vacuum" construction follows standard difference-spectrum practice in XAS and NEXAFS for isolating environment contributions when a true background is not directly measurable [89].

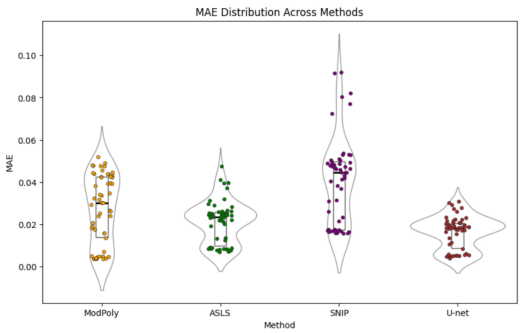

Figure 5.2.3: MAE on experimental test set of 54 $Ti\ L_{2,3}$ spectra of U-Net against classical methods.

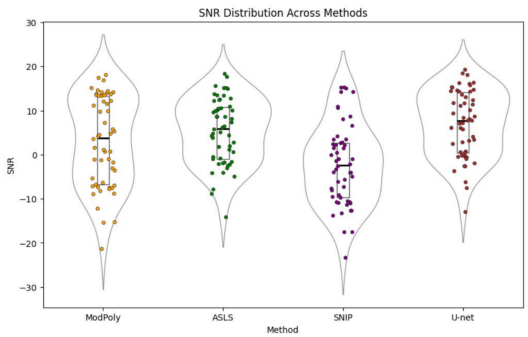

Figure 5.2.4: SNR on experimental test set of 54 $Ti\ L_{2,3}$ spectra of U-Net against classical methods.

The reported metrics in Table 5.2.1 represent mean across the 54 experimental test spectra. The detailed performance distributions are visualized in Figure 5.2.3 and Figure 5.2.4. Figure 5.2.3 compares the MAE across methods. The U-Net model

Table 5.2.1: Performance Metrics for Baseline Correction Methods and U-Net

Method	MSE	MAE	SNR (dB)
ModPoly	0.0053	0.0384	1.81
AsLS	0.0028	0.0269	3.93
SNIP	0.0020	0.0190	5.60
U-Net	0.0022	0.0166	6.42

demonstrated superior performance on experimental data, achieving the lowest MAE (0.0166) and the highest SNR (6.42 dB), with improvements of 12.6% in MAE and 14.6% in SNR over the next-best method (SNIP). The U-Net also exhibited lower variance in its predictions compared to most classical methods, indicating more consistent performance across diverse spectral morphologies.

5.3 Bayesian MCMC Approach: Rigorous Uncertainty Quantification

This section presents the second, complementary framework, a Bayesian inference method that prioritizes statistical rigor and physical interpretability over processing speed.

5.3.1 The Need For Uncertainty and Interpretability

Modern dDL pipelines for spectral analysis are attractive for their speed and automation, but produce deterministic point estimates with no calibrated error bars and limited physical interpretability. For XAS spectroscopy, where small shifts in edge energy, subtle near-edge features, and other effects can determine the scientific conclusion, results require explicit uncertainty quantification and a connection to physically meaningful parameters.

Rigorous statistical inference is therefore essential for spectroscopy analysis, particularly as models must account for increasing complexity. This complexity includes confounding factors such as instrument response, calibration drift, noise, feature overlap, among others. Capable inference methods are required to discern these factors, mitigate bias, and avoid overconfident parameter estimates. Furthermore, robust inference is critical for several practical reasons: i) analyses focus on low signal-to-noise feature, where uncertainty quantification is necessary to objectively distinguish spectral features from statistical fluctuations; ii) uncertainty quantification and model validation are required to prevent downstream analyses or theoretical modeling from being skewed by artifacts; iii) limited beamtime, sample degradation and other resource constraints make uncertainty framework important tools to guide decision

making of experiments; iv) uncertainty modeling are crucial when further data acquisition is impossible.
Bayesian inference directly addresses these requirements by conditioning a physically-motivated forward model on the experimental data. This methodology allows for the incorporation of all relevant information, including the complex noise properties and physical constraints, by separating the problem into two distinct components. The first one is a likelihood function, which defines the probability of the observed data given a set of model parameters. The second is a prior distribution, that specifies the initial knowledge about those parameters. The inference process then combines prior knowledge with the evidence from the data to compute the posterior probability distribution for every parameter. This distribution provides the comprehensive and rigorous uncertainty quantification required, trading the inference speed of the U-Net for full statistical interpretability, as will be detailed in the following section.

5.3.2 The Bayesian Inference Framework

The Bayesian inference framework approaches parameter estimation through the construction of a generative probabilistic model. This process requires the explicit formulation of all model components and assumptions, which are formalized in two core elements.
First, the likelihood function, $p(I_{\text{meas}}|\theta, I)$, quantitatively describes the probability of observing the measured data (I_{meas}) given a specific set of model parameters (θ). It encodes the core physical assumptions and noise characteristics of the measurement. For example, the functional form of the spectral features (e.g., peak shapes, arctan steps) and the stochastic model for the measurement (e.g. Gaussian noise) are captured within the likelihood's formulation.
Second, the prior probability distribution, $p(\theta|I)$, represents all prior knowledge or assumptions about the parameters before the data are considered. The choice of priors is a critical step in the model formulation. Priors may be uninformative (e.g., a broad uniform distribution) to allow the data to dominate the inference, or informative (e.g., a Gaussian distribution from a previous experiment) to incorporate existing knowledge. For physical models, priors are essential for enforcing known constraints, such as positive peak amplitudes and widths, or constraining edge positions to a physically plausible range.
Bayesian inference then uses Bayes' theorem as the mathematical rule for updating the prior beliefs with the evidence from the data:

$$p(\theta|I_{\text{meas}}, I) = \frac{p(I_{\text{meas}}|\theta, I) \cdot p(\theta|I)}{p(I_{\text{meas}}|I)} \tag{5.5}$$

Here, θ is the vector of all model parameters (e.g., peak positions, widths, ampli-

tudes, and background coefficients), I_{meas} is the observed NEXAFS spectrum, and I represents the prior information. The result is the posterior probability distribution, $p(\theta|I_{\text{meas}}, I)$. The denominator, $p(I_{\text{meas}}|I)$, is the marginal likelihood and serves as a normalization constant. For parameter estimation, the relationship is commonly expressed as a proportionality:

$$\underbrace{p(\theta|I_{\text{meas}}, I)}_{\text{Posterior}} \propto \underbrace{p(I_{\text{meas}}|\theta, I)}_{\text{Likelihood}} \cdot \underbrace{p(\theta|I)}_{\text{Prior}}$$

The posterior distribution constitutes the full solution to the inference problem and serves as the principal mechanism for quantifying uncertainty. Rather than providing a single best-fit estimate, it defines a multidimensional probability distribution over all plausible parameter values and their correlations. This formulation deals with uncertainty by allowing for marginalization over nuisance parameters (e.g., background coefficients, noise model parameters) to obtain the marginal posterior distribution for the specific parameters of scientific interest (e.g., a peak position). The framework delivers a statistically rigorous and fully calibrated estimate. In practice, because the posterior is rarely available in analytical form, it is typically explored numerically using sampling techniques such as MCMC. MCMC constructs a Markov chain whose equilibrium distribution converges to the target posterior, efficiently exploring potentially high-dimensional parameter spaces through iterative stochastic updates.

5.3.3 MCMC Implementation and Validation

The MCMC framework was implemented and then validated using two complementary studies: synthetic spectra with known ground truth parameters to enable quantitative accuracy assessment, and experimental Titanium spectra introduced in Section 5.1 to demonstrate real-world applicability.

The synthetic validation employed generated NEXAFS spectra with known ground truth parameters, allowing rigorous quantification of parameter recovery accuracy and uncertainty estimation reliability.

The forward model for synthetic spectra comprises three components: three absorption peaks, one arctangent absorption edge, and a cubic polynomial background $P_3(E) = a + bE + cE^2 + dE^3$:

$$I_{\text{model}}(E;\theta) = \sum_{i=1}^{3} I_{\text{peak},i}(E;\theta_{\text{peak},i}) + I_{\text{arctan}}(E;\theta_{\text{arctan}}) + P_3(E;\theta_{\text{bg}}) \tag{5.6}$$

The 13 free parameters consist of 6 peak parameters (width and height for each peak), 3 arctangent parameters (position, intensity, scale), and 4 polynomial background coefficients. Peak positions and profiles (1 Voigt, 2 Lorentzian) are fixed, assuming high SNR enables reliable peak localization via classical methods.

Measurement noise is modeled as additive Gaussian white noise with SNRs between 40 and 60 dB. The log-likelihood function is:

$$\log p(I_{\text{meas}}|\theta) = -\frac{1}{2\sigma^2}\sum_{j=1}^{M}[I_{\text{meas}}(E_j) - I_{\text{model}}(E_j;\theta)]^2 - M\log(\sqrt{2\pi}\sigma) \tag{5.7}$$

where M is the number of energy points and σ is the noise standard deviation derived from the specified SNR.

The MCMC sampling is performed using the emcee Python package [103], which implements an affine-invariant ensemble sampler. The sampler configuration employs 64 walkers and 10^5 total steps (including burn-in) for synthetic validation. Convergence is assessed by monitoring the integrated autocorrelation time. Sampling requires $\mathcal{O}(10)$ minutes on a single core of an AMD EPYC Genoa High Performance Computing (HPC) node. The synthetic spectra are generated with SNRs between 40-60 dB, representative of high-quality synchrotron measurements. Peak profiles are configured as one Voigt peak and two Lorentzian peaks, with fixed positions and gamma parameters while heights and widths remain free.

The synthetic validation demonstrates excellent parameter recovery across all model components, with peak heights and widths recovered within the statistical uncertainty of the input spectrum, as shown by the bottom panel of Figure 5.3.1.

Figure 5.3.1 shows the MCMC fit quality for the synthetic spectrum, demonstrating excellent agreement between forward model and data with minimal structured residuals.

This synthetic validation demonstrates that the MCMC framework successfully navigates the 13-dimensional parameter space and provides meaningful uncertainty quantification. The tight credible intervals (typically 1-5% of parameter values for peaks, 5-10% for background) establish confidence in applying the method to experimental data where ground truth is unavailable.

Following synthetic validation, the MCMC framework was applied to experimental NEXAFS spectra of titanium acquired at the APE-HE beamline. The sampler employs 64 walkers with 5×10^5 total steps for experimental data, requiring at most 1 hour on a single core of an AMD EPYC Genoa HPC node. The forward model is adapted to accommodate titanium's characteristic double-edge structure.

The titanium forward model extends the synthetic framework to accommodate the double-edge structure:

$$I_{\text{model}}(E;\theta) = I_{\text{bg}} + \sum_{i=1}^{6} I_{\text{peak},i}(E;\theta_{\text{peak},i}) + I_{\text{arctan},1}(E;\theta_1) + I_{\text{arctan},2}(E;\theta_2) \tag{5.8}$$

The 18 free parameters comprise 12 peak parameters (width and height for 6 peaks

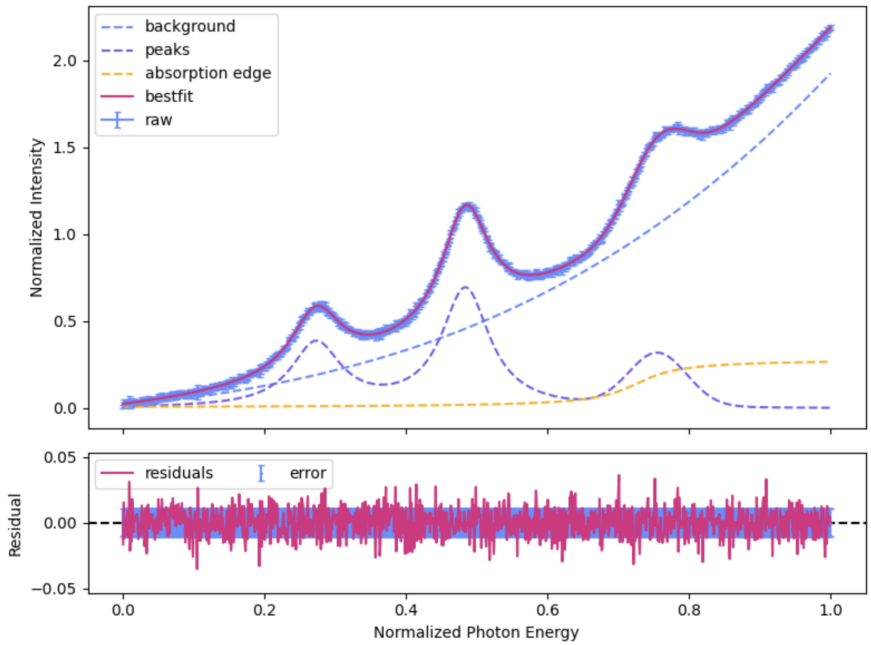

Figure 5.3.1: MCMC fit to synthetic NEXAFS spectrum with SNR ∼50 dB. The forward model (red) accurately captures all spectral components including three peaks, arctangent edge, and polynomial background. Residuals (bottom panel) show minimal structure, validating the forward model adequacy and parameter recovery.

modeling multiplet structure), 5 arctangent parameters (2 edges with constrained scale ratio reflecting $2p$ degeneracy), and 1 constant background. Peak positions and profiles remain fixed based on established Ti L-edge spectroscopy.

Uncertainty is estimated from two spectra acquired from the same experiment, with measurement uncertainties varying from $\sim 3 \times 10^{-4}$ to $\sim 3 \times 10^{-2}$ across the 363 energy points. This heteroscedastic noise model better reflects the experimental reality where SNR varies with spectral intensity, particularly across absorption edges. Errors remain assumed uncorrelated across energy points, though with position-dependent variance. Two independent titanium vacuum spectra serve as an ensemble, and parameter uncertainties are estimated from the mean and variance of fitted parameters across these measurements, providing an empirical assessment of reproducibility under nominally identical experimental conditions. The MCMC fit to the titanium data achieved a fit with reduced chi-squared of $\chi^2_{\text{red}} = 0.38$, visualized in Figure 5.3.2, which demonstrates the model's ability to accurately capture the complex multiplet structure and the double absorption edge. The residuals, plotted in the bottom panel, show minimal structure, confirming the adequacy of the forward model.

The primary advantage of the MCMC approach is the delivery of full posterior probability distributions for all model parameters, enabling statistically grounded credible intervals for derived physical quantities. This capability is essential for rigorous scientific interpretation and propagation of uncertainties in subsequent analyses.

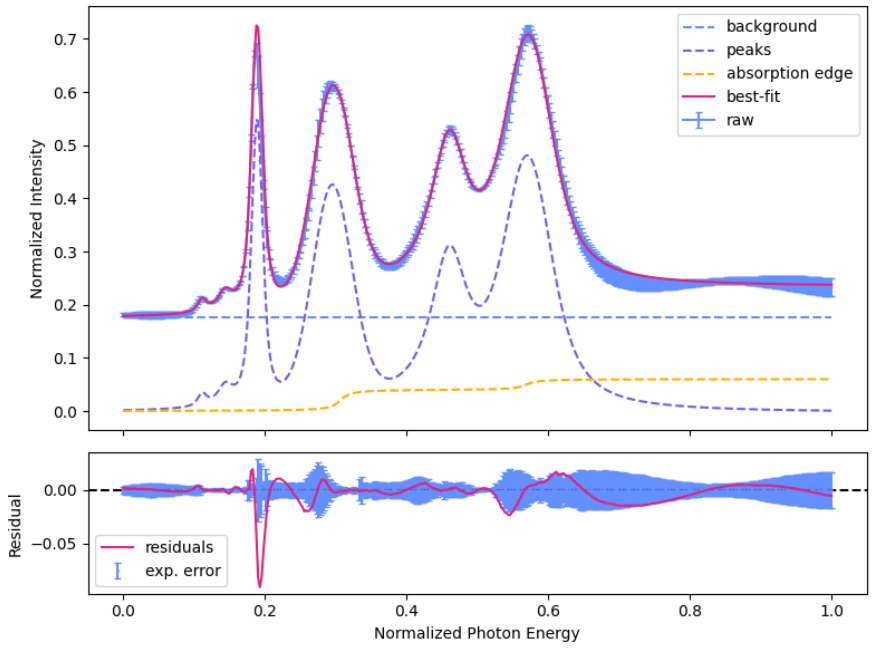

Figure 5.3.2: MCMC fit to experimental titanium $L_{2,3}$ spectrum. The forward model (red) accurately captures the multiplet structure, absorption edges, and background. Residuals (bottom panel) show minimal structure, indicating good model adequacy.

5.4 Discussion

This section compares the two proposed computational frameworks, discussing their respective limitations and outlining potential future research directions.

5.4.1 Method Comparison and Limitations

The two computational frameworks presented in this chapter address complementary objectives in NEXAFS spectroscopy analysis, with distinct advantages, limitations, and appropriate use cases. This section synthesizes their relative strengths and identifies methodological constraints that inform deployment strategies.

The U-Net framework addresses high throughput spectroscopy, enabling real time analysis compatible with synchrotron data acquisition rates. Full automation eliminates manual parameter tuning, ensuring reproducible results for identical inputs and removing operator-dependent variability. These characteristics make the U-Net approach well-suited for automated screening and operando reaction monitoring where rapid feedback guides experimental decisions.

However, fundamental limitations constrain appropriate application domains. The data-driven NN architecture functions as a "black box," offering no direct physical interpretability or mechanistic insight into spectral features. Point estimates provide no uncertainty quantification, precluding rigorous error propagation in derived quantities. Performance depends critically on training data distribution, with potential

degradation when applied to spectra exhibiting morphologies absent from synthetic training examples. All validation in this thesis employed titanium L-edge spectra from the APE-HE beamline, and generalization to substantially different materials, energy ranges, or experimental configurations requires empirical verification.

The MCMC framework provides rigorous statistical inference through explicit physics-based forward modeling. Full posterior probability distributions for all model parameters yield credible intervals that enable statistically grounded uncertainty propagation. The explicit forward model structure ensures interpretability, allowing direct connection between fitted parameters and physical properties such as peak positions, intensities, and widths. The framework naturally accommodates complex prior information, incorporating domain knowledge to constrain parameter spaces and improve inference in data-limited regimes.

However, several important limitations must be acknowledged. Computational cost remains substantial, requiring at most 1 hour per spectrum on a single core of an AMD EPYC Genoa HPC node (64 walkers, 5×10^5 steps for experimental data), limiting applicability to high-throughput screening scenarios. Reliable uncertainty quantification depends critically on accurate characterization of measurement errors. The current implementation assumes Gaussian white noise (constant or point-wise variance), implying uncorrelated measurement errors across energy points. This assumption may not hold when systematic drifts in beam intensity, detector response nonlinearity, or time-correlated electronic noise introduce correlations. When errors exhibit correlations, accurate uncertainty quantification requires construction of full covariance matrices, necessitating substantially larger measurement ensembles or sophisticated simulation-based error modeling.

The consequences of noise model limitations are evident in experimental fits, where $\chi^2 \ll 1$ indicates that measurement uncertainties were overestimated or that the assumption of uncorrelated errors fails to capture the true systematic structure of the data. Future work will warrant more sophisticated noise models or correlated error structures.

Model component choices such as peak profiles and background functional forms are fixed based on informed physical assumptions but may not hold universally across diverse chemical environments and experimental conditions. Peak profile shapes depend on complex combinations of instrumental resolution, core-hole lifetime broadening, and inhomogeneous effects whose relative contributions vary across energy ranges. Background shapes emerge from experimental conditions including temperature, pressure, gas composition, and sample geometry rather than being prescribed by theory. Fixed choices risk systematic errors when assumptions are violated.

The contrasting strengths motivate a workflow that leverages both methods strategically. The U-Net serves as a screening tool for high-throughput requirements or preliminary surveys characterizing spectral trends across sample series. Optional

out-of-distribution detection mechanisms can flag spectra exhibiting morphologies outside the training distribution, triggering escalation to more rigorous methods. The MCMC approach is then selectively applied to critical measurements where computational cost is justified: analysis requiring uncertainty propagation, detailed physical interpretation of peak structure and edge positions, or validation of U-Net predictions on novel spectral morphologies. This strategy balances the throughput demands of modern synchrotron facilities with the statistical rigor expected in quantitative scientific analysis.

5.4.2 Future Directions

Several methodological extensions would enhance both frameworks and address current limitations.

Accurate uncertainty quantification requires realistic error models that capture the true noise structure in experimental data. Future work should develop empirical noise characterization protocols through systematic measurement repetition under controlled conditions, estimating both noise variance and correlation structure. Simulation campaigns could model beamline response, detector characteristics, and environmental effects to construct realistic covariance matrices. Such sophisticated error treatment would improve credible interval reliability, particularly for derived quantities depending on multiple correlated parameters.

The Akaike Information Criterion (AIC) provides a principled framework for model selection that balances goodness of fit against model complexity. The AIC is defined as:

$$\text{AIC} = 2k - 2\ln(\hat{L}) \tag{5.9}$$

where k is the number of free parameters and $\hat{L}$ is the maximum likelihood value achieved by the model. Lower AIC values indicate preferable models, with the $2k$ penalty term naturally discouraging overfitting. By computing AIC values for competing models (e.g., Gaussian vs. Voigt peak profiles, polynomial vs. exponential backgrounds), the most appropriate model structure can be selected objectively rather than by assumption. Systematic exploration across model spaces would identify optimal forward model configurations for diverse chemical systems and experimental conditions.

Both frameworks require validation beyond titanium L-edge spectra from the APE-HE beamline. Testing on transition metal K-edges, rare earth M-edges, and light element K-edges would assess generalization across chemical environments and energy ranges. Cross-beamline validation at facilities with different instrumental resolutions, energy calibrations, and detector technologies would establish robustness to experimental variability. Such comprehensive validation would define confidence bound-

aries for method applicability and identify systematic limitations requiring methodological refinement.

Emerging methods combining DL with physics-based interpretability offer promising directions. Physics-informed NNs incorporate forward model constraints directly into loss functions, potentially improving generalization while retaining interpretability. Exploration of such hybrid architectures could yield methods that balance the speed advantages of DL with the uncertainty quantification and interpretability of Bayesian inference.

Operational deployment requires integration with existing data management infrastructure at synchrotron facilities. Real-time U-Net processing could provide immediate feedback during data acquisition, enabling adaptive experimental strategies that respond to preliminary analysis results. Automated quality control metrics could detect instrumental anomalies or sample degradation, triggering operator alerts. MCMC analysis could be dispatched to HPC resources as background jobs, with results delivered asynchronously once sampling converges. Such integrated workflows would transition these methods from research prototypes to operational tools supporting routine facility operation.

5.5 Conclusions

This chapter demonstrates that complementary computational strategies can address competing demands in scientific data analysis. The DL U-Net approach prioritizes speed and automation for high-throughput screening, while the Bayesian MCMC framework emphasizes rigorous uncertainty quantification and physical interpretability. Neither method dominates universally; rather, their optimal deployment depends on specific analytical objectives, available computational resources, and required statistical rigor. The proposed workflow architecture combines automated screening with selective rigorous analysis, offering a generalizable framework for spectroscopic modalities. More broadly, this thesis illustrates how physics-informed ML and Bayesian inference can be strategically combined to balance the practical demands of modern experimental facilities with the statistical standards expected in quantitative science.

Chapter 6

Deployment and Integration of services

The research contributions described in Chapters 2 to 5 address challenges in nanoscience data management and analysis. This chapter documents the deployment of these methods from research prototypes to operational services within Italian and European research infrastructures.
Two deployment contexts are described: i) production web services, collectively named TriDAS, operating within the NFFA-Europe project; and ii) interactive computational notebooks developed for the NFFA-DI project. These deployments validate the practical applicability of the FAIR-by-design approaches established in earlier chapters and demonstrate their integration with Italian and European nanoscience facilities.
The rest of the chapter is organized as follows: Section 6.1 summarizes NFFA-Europe and NFFA-DI projects, AREA Science Park computational resources and data infrastructure; Section 6.2 describes the production web services within NFFA-Europe; Section 6.3 details the Jupyter notebook-based services on NFFA-DI digital infrastructure; Section 6.4 discusses future service development and the planned integration of services between the projects; Section 6.5 concludes the chapter.

6.1 European Research Projects and AREA Infrastructure

The deployment of these services is embedded within the context of existing European and Italian research infrastructures, which provide the necessary computational resources and data ecosystems.

6.1.1 NFFA-Europe Pilot and NFFA-DI Context

NFFA-Europe is a Horizon 2020 research infrastructure project coordinated by CNR-IOM[104]. The project establishes an Interoperable Distributed Research Infrastructure for Nanoscience (IDRIN) spanning multiple European facilities. NFFA-Europe provides integrated ac-

cess to nanoscience capabilities, including materials synthesis, nanofabrication, characterization, and numerical simulation resources.

NFFA-Europe combines traditional Transnational Access (e.g. physical access to facilities) with a new modality called Virtual Access (VA). VA provides online services for data exploration, analysis, and processing, enabling researchers to interact with experimental data and computational tools without requiring physical presence at facilities. This dual-access model addresses both hands-on experimental work and the need for remote data analysis.

Data generated at Nanoscience Foundries and Fine Analysis (NFFA) facilities is managed according to FAIR guidelines, with integration into the European Open Science Cloud (EOSC) ecosystem [105].

NFFA-DI is a parallel Italian initiative funded through Italian National Recovery and Resilience Plan (PNRR) coordinated by CNR-IOM[106]. NFFA-DI focuses on upgrading the Italian nanoscience infrastructure by integrating nanofoundry laboratories, structural characterization facilities, Elettra synchrotron radiation facilities, and technology transfer pathways to intermediate Technology Readiness Levels (TRL). The infrastructure aims to bridge the gap between fundamental quantum matter research and functional micro-systems for digital transformation applications.

NFFA-DI's digital strategy emphasizes a single-entry user portal, service catalogue, FAIR data infrastructure, and technology transfer support. The investment in digital infrastructure complements the physical facility upgrades, ensuring that data management and analysis capabilities keep pace with advanced instrumentation.

NFFA-Europe is part of the European Strategy Forum on Research Infrastructures (ESFRI) roadmap [107], connecting to other materials science, photon and neutron, and analytical infrastructure initiatives. This European coordination enables cross-infrastructure data sharing, standardized metadata schemas, and collaborative service development. The scale of these initiatives provides context for the deployment of FAIR-by-design services as part of a broader transformation in how European nanoscience research is conducted, shared, and reused.

Figure 6.1.1 illustrates the consortium structure of NFFA-Europe and NFFA-DI, showing participating institutions across European and Italian facilities. The distributed nature of these infrastructures requires a robust digital infrastructure for data sharing and remote service access, providing context for TriDAS and NFFA-DI VA deployments described in Sections 6.2 and 6.3.

The VA services developed in this thesis operate within this infrastructure: TriDAS services (Section 6.2) represent a subset of NFFA-Europe VA offerings, while Jupyter notebooks [108] (Section 6.3) constitute the NFFA-DI VA component. Both service portfolios translate the research contributions from Chapters 2 to 5 into operational tools serving the European nanoscience community.

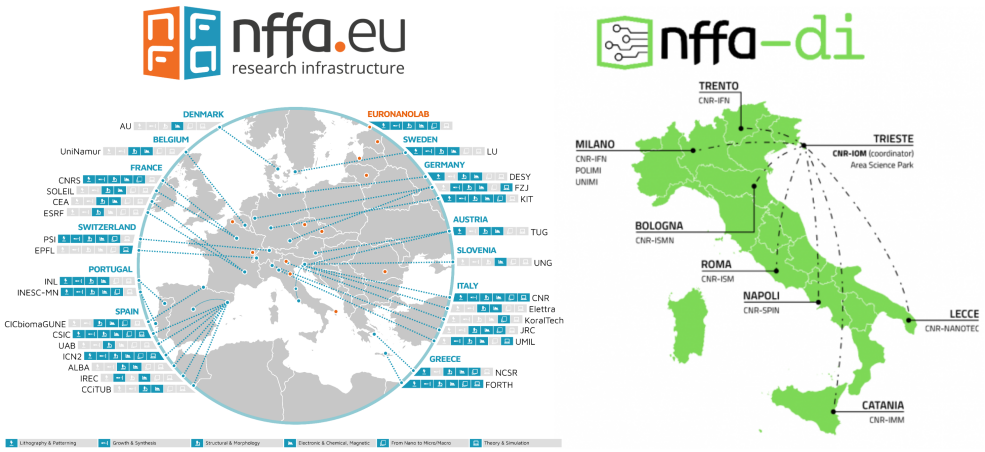

Figure 6.1.1: NFFA-Europe and NFFA-DI consortium partners showing distributed infrastructure across European and Italian facilities.

6.1.2 ORFEO Datacenter Infrastructure

The ORFEO datacenter [109], located at AREA Science Park in Trieste, Italy, provides the computational and storage infrastructure supporting the digital infrastructure of NFFA-DI and the TriDAS services of NFFA-Europe described in this thesis. ORFEO is a HPC and AI facility developed through regional and national investments between 2020 and 2025.
The facility integrates heterogeneous compute nodes interconnected via an Infiniband low-latency network. Current computational capabilities comprise:

- HPC Compute Servers: 33 nodes featuring Intel and AMD processors, totaling over 2,800 CPU cores, dedicated to simulations, numerical modeling, parallel computing, and large-scale data processing.
- AI/ML Compute Servers: 6 GPU nodes equipped with 48 NVIDIA accelerators, composed of 8 V100, 16 A100, and 24 H100 models, available for ML research workloads.
- Cloud Services Servers: 9 nodes hosting the ORFEO Kubernetes Cluster, providing the Infrastructure as a Service Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) environment for the containerized services described in this chapter.

ORFEO storage exceeds 7 PB, organized in performance tiers managed through a parallel Ceph filesystem. This architecture provides high speed storage for active workloads and high capacity archival storage.
ORFEO connects to research networks and utilizes, whenever possible, an open-standard software stack including Linux, Slurm [110], Ceph [111], and Kubernetes [112], optimized for scientific computing and AI. The Kubernetes cluster running

on the Cloud Services Servers provides the orchestration for TriDAS services (Section 6.2), while the Ceph filesystem provides persistent storage accessible by analysis services.

6.1.3 OFED Digital Ecosystem

The Overarching FAIR Ecosystem for Data (OFED) provides the FAIR data management infrastructure supporting NFFA-DI services. OFED integrates five components that enable data management from collection to publication. The overall workflow, illustrated in Figure 6.1.2, shows the progression from experimental data acquisition and FAIR data processing into the OFED environment for management, analysis, and eventual publication.

The primary component is Novel Materials Discovery (NOMAD) Oasis, an instance of the NOMAD [113] platform developed by the FAIRmat consortium [114]. NOMAD Oasis is a material science data repository, that includes metadata management, Electronic Laboratory Notebook (ELN), DOI assignment, and RESTful Application Programming Interface (API) integration. NOMAD Oasis enables NFFA-DI to maintain data sovereignty through local deployment while remaining interoperable with the FAIRmat ecosystem and the main NOMAD instances.

The second component is the Ceph distributed storage system [111] hosted at ORFEO (described in Section 6.1.2). Accessed via the Ceph object storage RADOS [115], it serves as the shared data layer between NOMAD Oasis and JupyterHub [116]. This shared architecture enables users to access their data from both the repository interface and computational notebooks without requiring data duplication.

The third component is Authentik, an open-source identity provider implementing Single Sign-On (SSO) protocols [117]. Authentik provides centralized authentication for OFED components, including NOMAD Oasis, JupyterHub, and also integrates with the TriDAS services, enabling researchers to access the complete ecosystem with unified credentials.

The fourth component is EasyDMP, a web application for creating a Data Management Plan (DMP) [118]. This standalone tool facilitates the planning phase of data management, enabling metadata schema definition and preservation strategies for laboratories, instruments, and research proposals.

The fifth component is JupyterHub, providing a computational environment for executing NFFA-DI VA Service notebooks (Section 6.3.1). JupyterHub shares the Ceph RADOS data layer with NOMAD Oasis, enabling users to access repository data from computational notebooks, perform analysis, and deposit results back to NOMAD Oasis with provenance tracking.

The OFED infrastructure supports a structured and interoperable data workflow, as shown in Figure 6.1.2. Data and metadata are collected from instruments and Elec-

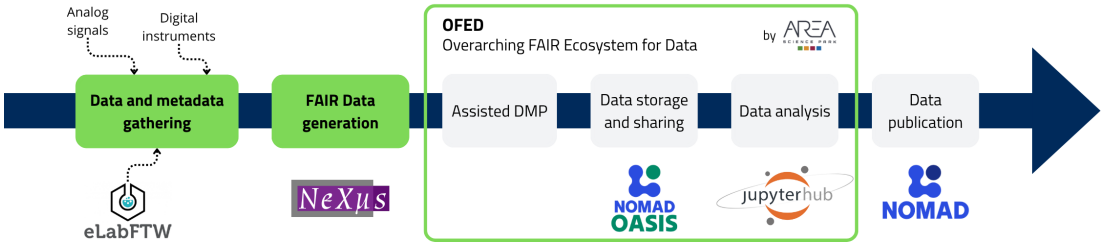

Figure 6.1.2: The diagram illustrates the data lifecycle progression: (i) data and metadata gathering from instruments and ELNs such as elabFTW; (ii) FAIR data generation, such as using NeXus format; (iii) data management planning and documentation, facilitated by tools like EasyDMP; (iv) ingestion into the OFED ecosystem, managed by NOMAD Oasis and stored within the underlying Ceph system; (v) access for data analysis via integrated services like JupyterHub; and (vi) potential publication in the central NOMAD instance.

tronic Lab Notebooks, such as elabFTW [119]. These inputs undergo FAIRification, typically using standardized formats like NeXus [28]. DMPs are created with the EasyDMP tool integrated into OFED, after which the FAIR data are ingested into NOMAD Oasis, which uses the underlying Ceph storage system. The data are then available for analysis through environments such as JupyterHub. Finally, curated datasets can be published through the central NOMAD repository for dissemination and DOI assignment. This workflow ensures the preservation of data provenance and adherence to FAIR principles throughout the data lifecycle.

6.2 TriDAS Services within ORFEO Kubernetes

As a contribution to the NFFA-Europe's VA portfolio, this thesis developed the TriDAS [120]. These production services implement the FAIR-by-design data management principles established in Chapter 2 and are deployed on the ORFEO Kubernetes infrastructure (Section 6.1.2).

6.2.1 High-Level Architecture

TriDAS services are deployed as containerized microservices on the ORFEO Kubernetes cluster, accessible at `https://TriDAS.nffa.eu/`. Figure 6.2.1 illustrates the architecture.

The Nginx Ingress Controller [121] handles incoming HyperText Transfer Protocol Secure (HTTPS) requests and routes them to the appropriate backend service based on Uniform Resource Locator (URL) path. Each service runs within the dedicated `area-TriDAS` namespace.

Persistent storage for the deployed datasets is provided via Kubernetes Persistent Volume Claims (PVC), backed by the ORFEO Ceph infrastructure (Section 6.1.2). These volumes are mounted into the service containers, ensuring data availability for:

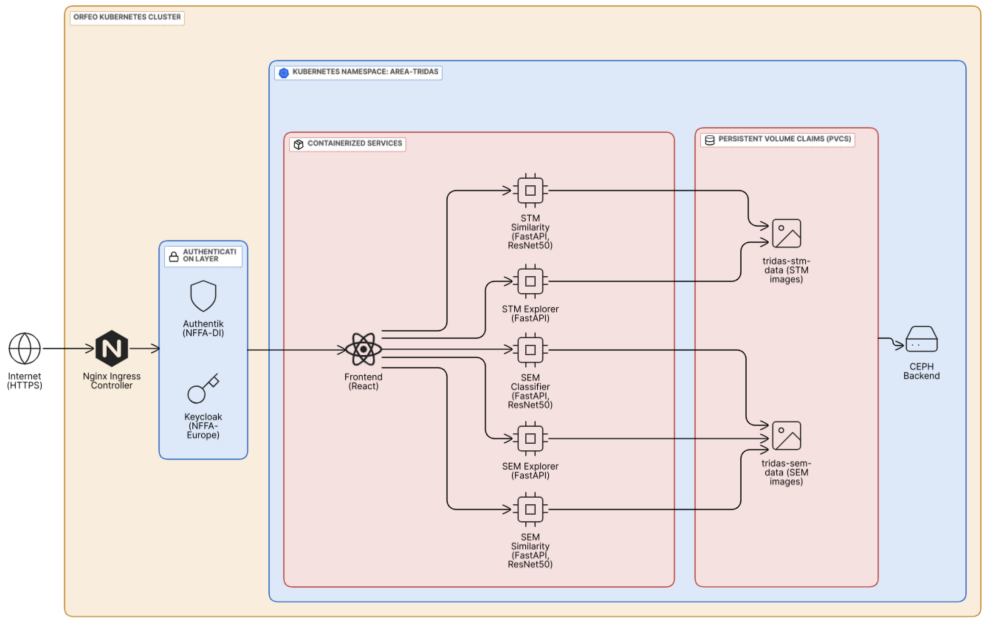

Figure 6.2.1: TriDAS Kubernetes Architecture showing Nginx Ingress Controller routing to containerized services in the `area-TriDAS` namespace, with persistent volume claims for STM and SEM datasets.

- The STM dataset comprising 7,287 images, which was fully curated and described in Section 2.4.
- The SEM Majority dataset [122], a published dataset of 25,537 SEM images. This dataset consists of images classified into ten predefined morphological categories: (1) Porous sponges, (2) Patterned surfaces, (3) Particles, (4) Films coated surfaces, (5) Powder, (6) Tips, (7) Nanowires, (8) Biological, (9) MEMS devices and electrodes, and (10) Fibres.

Authentication integrates with the NFFA-Europe Keycloak [123] SSO system. Users authenticate via Keycloak using their NFFA credentials, and a JSON Web Token (JWT) is used to authorize access to the TriDAS services. Furthermore, federation between Keycloak and the ORFEO Authentik identity provider (Section 6.1.3) allows users with NFFA-DI credentials seamless access.

6.2.2 Deployed Services

Five services are currently deployed on the TriDAS platform, each addressing specific data exploration and analysis needs for STM and SEM images.

STM Explorer. This service addresses the challenge of efficiently navigating large, heterogeneous experimental collections of STM images. STM Explorer provides interactive metadata filtering and visualization for the 7,287 image STM dataset curated

in Chapter 2. Researchers can query, sort, and visualize images using experimental metadata. A subset of 11 labels was carefully selected, that provide significant information about image characteristics and microscope settings, and is described in Table 6.2.1.

Metadata	**Description**
Date	image acquisition date (YYYY-MM-DD)
FieldXSizeinnm	X dimension of the scan size, in nanometers (nm)
FieldYSizeinnm	Y dimension of the scan size, in nanometers (nm)
XOffset	X coordinate of the tip offset from the center of the scan axes, in nanometers (nm)
YOffset	Y coordinate of the tip offset from the center of the scan axes, in nanometers (nm)
ScanSpeed	speed of the scan, in nanometers per second (nm/s)
ScanAngle	rotation angle of the fast scan direction in the XY plane, in degrees (°C)
GapVoltage	bias voltage applied between tip and sample in the constant current scan mode, in Volts (V)
LoopGain	integral term of the PID feedback loop controller of the tunneling current
FeedbackSet	setpoint of the tunneling current, in nanoamperes (nA)
Label	sample material composition (categorical: Gr_Ni100, Gr_Ni111, N_Gr_Ni111)

Table 6.2.1: STM Explorer metadata available.

This design allows for rapid identification of clusters, experimental outliers, and underlying parameter correlations within the dataset.
Figure 6.2.2 summarizes the workflow: i) users select metadata fields to generate plots, ii) inspect clusters through a metadata table, and iii) preview images corresponding to selected data points. In addition, images can be downloaded together with raw data and W3C-PROV provenance metadata. The service uses a FastAPI backend to efficiently query metadata and retrieve images, while a Bokeh Server frontend provides interactive visualizations. Metadata is indexed in memory for fast access, and images are served directly from persistent storage.

STM Similarity. This service addresses the need to discover morphological similarities and common visual patterns in STM images. STM Similarity implements CBIR for the STM dataset of Chapter 2. The service uses feature embeddings extracted by a pre-trained ResNet50 model to capture visual characteristics of STM images. Users upload a query image, and the service computes the cosine similarity between the query embedding and the embeddings of the dataset images, returning the most visually similar images ranked by similarity score.
This approach enables discovery of images with comparable visual patterns without requiring specific metadata knowledge, complementing services such as STM explorer.

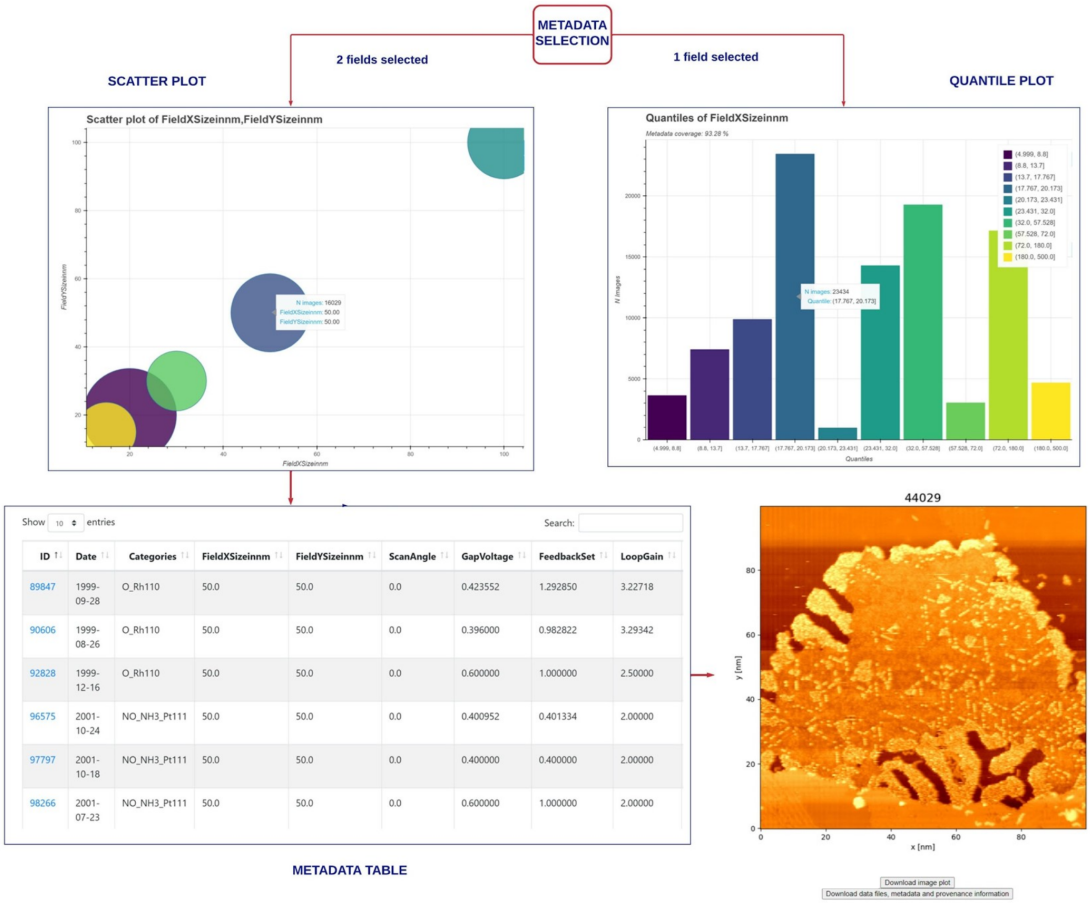

Figure 6.2.2: STM Explorer interface panels demonstrating the metadata exploration workflow, including metadata selection, plot visualization, table filtering, and image preview.

Feature embeddings are precomputed for all dataset images to ensure rapid query processing, only the query image embeddings are computed on invocation.
Results are presented as a ranked gallery displaying the retrieved images alongside their similarity scores. Downloaded result sets include both similarity scores and original metadata, facilitating analysis of how visual similarity relates to experimental conditions.

SEM Explorer. This service addresses the challenge of efficiently navigating large, heterogeneous experimental collections of SEM images. SEM Explorer provides interactive metadata filtering and visualization for the 25,537 images of the SEM Majority dataset. Its functionality, user workflow, and technical implementation are analogous to those described for STM Explorer. The key metadata fields available for filtering and visualization are detailed in Table 6.2.2. While derived from instrument-specific parameters, this schema can be extended towards community standards, such as the definitions provided by the Helmholtz Electron Microscopy Glossary [124], to ensure broader semantic interoperability.

Metadata	**Description**
Date	image acquisition date (YYYY-MM-DD)
Magnification	magnification level of the acquired image (e.g., 1000×, 10000×)
AcceleratingVoltage	accelerating voltage of the electron beam, in kilovolts (kV)
eBeamWorkingDistance	distance between the sample surface and the final lens, in millimeters (mm)
PixelSize	physical size of each pixel in the image, in nanometers (nm) or micrometers (μm)
Category	morphological classification (categorical: Porous Sponge, Patterned Surface, Particles, Films, Powder, Tips, Nanowires, Biological, MEMS, Fibres)

Table 6.2.2: SEM Explorer metadata available.

SEM Similarity. This service addresses the need to discover morphological similarities and common visual patterns in SEM images. SEM Similarity implements CBIR for the SEM Majority dataset, analogous to the STM Similarity service.

SEM Classifier. This service addresses the need for automated organization and annotation of large SEM image archives. SEM Classifier provides automated image classification for the SEM Majority dataset, assigning images to one of the ten predefined morphological categories. The classification is performed using a ResNet50 CNN model trained on this dataset, following previous work methodology [122]. Users can upload one or multiple SEM images, and the service returns the predicted category for each image along with a confidence score.

6.3 NFFA-DI Services within OFED

While TriDAS services focus on microscopy data exploration and analysis, the broader NFFA-DI ecosystem incorporates additional data management and analysis capabilities centered on the NOMAD repository and Jupyter notebook-based workflows. This section describes the NOMAD repository selection rationale, the NFFA-DI Jupyter notebook services, and the integrated data workflow spanning instrument acquisition through publication.

6.3.1 NFFA-DI Jupyter Notebook Services

This thesis developed six Jupyter notebook services for microscopy image analysis within the OFED JupyterHub environment. These notebooks access experimental datasets from NOMAD Oasis via the shared data layer and deposit analysis results back to the repository.
The critical distinction between NFFA-DI notebooks and TriDAS services lies in deployment modality: TriDAS services are production-deployed FastAPI web applications running on Kubernetes (Section 6.2), whereas NFFA-DI notebooks are interactive Jupyter environments executed within JupyterHub. This difference reflects the separate project contexts and provides complementary access patterns for researchers preferring interactive exploration via web service.
The six NFFA-DI notebook services are as follows:

Image Feature Extraction Service This service addresses the need to transform raw image data into compact numerical representations suitable for downstream ML tasks. It performs deep feature extraction using ViTr models from the HuggingFace Transformers library [125], generating high-dimensional feature embeddings (typically 768 or 1024 dimensional vectors) for downstream tasks including classification, clustering, and similarity analysis. The ViT architecture captures both local texture features and global structural patterns through multi-headed self-attention mechanisms, providing richer representations than the spatial features extracted by CNNs.

Feature Similarity Service This service addresses the need to discover morphological similarities and common visual patterns in images by using distance-based similarity retrieval using ViT feature embeddings. This service applies the same CBIR concept as TriDAS STM/SEM Similarity services (Section 6.2.2) but leverages ViT embeddings instead of ResNet50 features. Cosine similarity ranking identifies visually similar images from curated datasets, enabling researchers to find experimental conditions producing similar morphologies or structures.

SEM Categories Classification Service This service addresses the need for automated classification of SEM images in broad morphological categories. It provides automated 10-category SEM image classification using a fine-tuned ViT model. This service represents an evolution and upgrade of the TriDAS SEM Classifier (Section 6.2.2), which uses ResNet50 architecture. Both services classify images into the same 10 categories, but the ViT-based implementation provides improved classification accuracy through the transformer architecture's superior capability to capture long-range dependencies in images.

SEM Scale Classification Service The notebook addresses the challenge of unreliable scale information in large SEM archives, which is often hindered by proprietary file formats. It classifies images in discrete magnification levels using a ViT model fine-tuned on the SEM Majority dataset.

STM Artifact Classification Service This service addresses the need for automated quality control in high-throughput data acquisition of STM images by identifying artifacts. It implements binary classification of STM images (artifact-free versus multi-tip artifacts) using a fine-tuned ViT model. This service represents direct implementation of the multi-tip artifact detection methods developed in Chapter 3, where the combination of Fourier transform analysis and ViT classification achieved high accuracy in identifying problematic images. The deployment of this capability as a JupyterHub service demonstrates the transition of research methods to operational infrastructure serving the NFFA research community.

SEM OCR Metadata Extraction Service This service addresses the challenge of recovering experimental metadata embedded in images as text overlays or embedded in proprietary headers. It performs comprehensive metadata extraction by combining Tesseract Optical Character Recognition (OCR) with embedded Tagged Image File Format (TIFF) metadata parsing. This service addresses FAIR metadata enrichment for legacy SEM datasets where experimental parameters are recorded as text overlays on images rather than in structured metadata fields. The service extracts embedded TIFF tags, applies OCR to identify text regions and parse experimental parameters. The notebook is based on the sem-meta python package, developed for this task [126].

NFFA-DI services developed in this thesis follow open-science principles. Notebooks are released under the GNU Affero General Public License (GNU AGPL) v3.0, documentation is licensed under CC BY 4.0 [127]. Additionally, trained models are publicly available [128], hosted on HuggingFace [129] under Apache 2.0 license.

6.4 Future Service Integration

The primary deployment objective is to provide an implementation as Jupyter notebooks within the OFED JupyterHub, which enables validation and iterative refinement with domain experts. Integration with NOMAD Oasis through the shared Ceph data layer ensures that processed data maintains provenance to original experimental acquisitions, preserving FAIR principles throughout the workflow.
The image restoration methods (Chapter 4) and spectroscopy background removal techniques (Chapter 5) represent logical extensions to the service portfolio described in this chapter. Both methods address preprocessing requirements common across NFFA facilities: the U-Net architecture for NEXAFS background removal can generalize to other spectroscopy modalities including X-ray Photoelectron Spectroscopy (XPS), Raman, and infrared spectroscopy, while the diffusion-based image restoration approach for STM can potentially be extended to other microscopy techniques. The MCMC method (Chapter 5) requires significant computation, making it less suitable for an interactive notebook service. It may, however, be provided for specific high-performance workflows.
Additionally, the integration of NFFA-DI notebooks as services within TriDAS is planned following completion of this thesis.

6.5 Conclusions

This chapter documented the deployment of FAIR-by-design ML methods developed within European research infrastructure. Two deployment contexts were illustrated: production web services operating as TriDAS within NFFA-Europe, and interactive Jupyter notebook services developed for NFFA-DI.
The TriDAS platform provides five operational services for STM and SEM data exploration, similarity search, and classification, deployed on ORFEO Kubernetes.
The NFFA-DI Jupyter notebook services offer a complementary portfolio of six ViT-based analysis workflows. The direct implementation of multi-tip artifact detection (Chapter 3) as an operational service exemplifies the transition from research contribution to infrastructure capability.
The services described demonstrate that FAIR-by-design approaches can scale from research prototypes to production infrastructure serving distributed European facilities. The notebook-to-service conversion pathway established through NFFA-DI provides a sustainable mechanism for continued development, while the cloud-native architecture of TriDAS ensures portability and long-term maintainability. Future integration of image restoration (Chapter 4) and spectroscopy background removal (Chapter 5) capabilities will further extend the service portfolios of NFFA-DI and NFFA-Europe projects.

Chapter 7

Conclusions

This thesis demonstrates a methodology for complete workflow development in AI-powered nanoscience research, progressing from FAIR data curation through automated quality control and computational enhancement to production service deployment within European and Italian research infrastructures. The thesis bridges fundamental challenges in materials characterization—unstructured legacy data, experimental artifacts, slow acquisition, manual analysis bottlenecks—with operational AI services accessible to the European and Italian research communities through NFFA-Europe and NFFA-DI platforms. In doing so, this thesis confirmes the central research hypothesis that a FAIR-by-design approach provides a viable methodological path to integrate AI across experimental techniques, transforming curated data into reliable, operational services for the nanoscience community.

The integrated approach addresses five interconnected objectives spanning foundational data management (Chapter 2), automated quality control (Chapter 3), computational enhancement (Chapter 4), spectroscopy analysis automation (Chapter 5), and sustainable deployment (Chapter 6). Three overarching methodological themes unify these contributions: FAIR-by-design methodology embedding findability, accessibility, interoperability, and reusability from research inception; physics-informed ML addressing experimental data scarcity through synthetic generation from accurate forward models; and ViT architectures demonstrating substantial advantages over established CNNs for specialized microscopy tasks.

7.1 Contributions to FAIR-by-Design Path

The following subsections directly address the research questions posed in Chapter 1. They detail the contributions and outcomes that answer each question, demonstrating how the thesis objectives were achieved.

7.1.1 FAIR Data Foundation

Chapter 2 developed automated workflows that enable FAIR-compliant publication of large experimental STM datasets while minimizing manual annotation burden. Automated metadata extraction processed instrument parameter fields from proprietary binary formats, establishing the foundation for systematic dataset organization. Semi-automated labeling combined DL for image similarity retrieval with expert verification. This hybrid approach significantly reduced manual annotation effort while maintaining quality control. W3C-PROV provenance documented the complete data history, transforming a large raw archive into a reference dataset and a subsequently refined, curated dataset with verified material category labels. Both datasets were published on Zenodo and were assigned persistent DOIs and permissive licensing, with a F-UJI FAIR assessment confirming compliance. The resulting datasets serve as the foundational data for subsequent quality control methods and deployed services, demonstrating the practical value of data stewardship.

7.1.2 Automated Quality Control

To address automated quality control, Chapter 3 developed FFT-enhanced ViT architectures for automated multi-tip artifact detection despite severe class imbalance. A three-channel input representation combined real and FFT space, enabling NNs to exploit both spatial and frequency domain features characteristic of artifact duplication patterns. The proposed ViT architecture achieved high detection accuracy and substantially outperformed traditional CNN baselines. Ablation studies confirmed that frequency-domain information, particularly the FFT amplitude, was the dominant discriminative feature. This thesis demonstrate ViT effectiveness for specialized microscopy, motivating the architectural evolution in the NFFA-DI services.

7.1.3 Physics-Informed Image Enhancement

Chapter 4 developed generative models enabling STM image restoration and super-resolution through physics-informed synthetic data generation, addressing data scarcity challenges inherent to specialized experimental techniques. Realistic degradation models captured multi-tip artifacts, scan-line noise, tip blurring, and sudden tip-change events. The FM generative model achieved high-fidelity artifact removal, outperforming both DDIM models and the autoencoder baseline on several validation metrics. The method demonstrated super-resolution capabilities while maintaining structural fidelity, with inference times practical for deployment on consumer-grade hardware. This restoration methodology provides complementary capabilities to the artifact detection from Chapter 3.

7.1.4 Spectroscopy Background Removal

Chapter 5 presented complementary approaches to NEXAFS spectroscopy background removal. A large, physics-informed synthetic dataset was generated to train the U-Net, eliminating the need for manual background annotation. Benchmarks demonstrated the U-Net outperformed classical algorithms on key performance metrics. In contrast, a Bayesian MCMC sampler provided full posterior distributions for rigorous uncertainty quantification, although at a significantly higher computational cost. The U-Net's speed advantage over MCMC enables real-time high-throughput processing, while the MCMC approach is reserved for rigorous analysis where statistical confidence bounds are essential.

7.1.5 Deployment to European Infrastructure

Chapter 6 documented production service deployment within the European and Italian nanoscience infrastructure, demonstrating sustainable pathways from research prototypes to accessible services. Five TriDAS web services were deployed on ORFEO Kubernetes infrastructure (2021-2023) using ResNet50 architectures, providing RESTful APIs and JavaScript frontend applications.
Six NFFA-DI Jupyter notebook services were developed (2024-2025) incorporating ViT architectures based on Chapter 3 findings, including a direct implementation of the multi-tip artifact detection method. The deployed infrastructure serves the research community through trans-national access programs. This evolution from the initial implementation of TriDAS services to the one used in NFFA-DI services documents a realistic deployment dynamic, where production stability requirements create an expected lag in adopting the latest research architectures.

7.2 Limitations and Future Directions

Rigorous assessment requires honest acknowledgment of limitations alongside contributions. This section consolidates key constraints encountered and outlines primary future research directions addressing identified gaps.

7.2.1 Key Limitations

Dataset scope constrains generalization claims across the thesis. The STM datasets presented in Chapters 2 and 3 originated from a single laboratory using consistent instrumentation over twenty years, providing temporal depth but limiting instrumental diversity, which is also the case for Chapter 4. The NEXAFS analysis in Chapter 5 focused on single material system, absorption edge, and single beamline, requiring future validation across diverse material classes, different edges, and multiple synchrotron facilities.

Workflow integration gaps separate independently developed capabilities from unified operational systems. Chapter 3 detection and Chapter 4 restoration operate standalone without orchestrated pipelines combining automated screening, quality validation, and expert review interfaces. Full integration requires workflow orchestration infrastructure, user interfaces supporting human oversight, and provenance tracking distinguishing automated versus manual decisions.

Architecture deployment lag affects production service performance. TriDAS services (deployed 2021-2023) employ ResNet50 architectures despite newer architectures advantages such as ViT. This lag reflects production infrastructure stability requirements, though planned TriDAS upgrades will address this gap while maintaining backward API compatibility.

Ground truth ambiguity introduces evaluation uncertainty across multiple chapters. Chapter 3 multi-tip detection relies on manual expert annotation potentially exhibiting inter-annotator variability for subtle artifacts. Chapter 4 restoration validation assumes physics-informed forward models accurately represent experimental degradation, with synthetic-real distribution mismatch potentially causing overly optimistic performance estimates. Chapter 5 NEXAFS background removal lacks objectively true backgrounds, as different analysts may fit different background functions yielding distinct results—automated methods provide consistency advantages but cannot claim absolute correctness when ground truth itself proves ambiguous.

7.2.2 Primary Future Directions

Integration of detection and restoration workflows represents a priority, which involves implementing automated pipelines that combine multi-tip classification and image restoration and super resolution. The required components include workflow orchestration to manage task dependencies, automated quality validation, user interfaces to support expert review, and W3C-PROV provenance tracking.

Cross-material and cross-facility validation is essential to establish the generalization bounds of these methods. The workflows detailed in Chapters 2 to 4 require testing on STM datasets from other laboratories with different instruments, operator practices, and material systems. Similarly, the spectroscopy methods from Chapter 5 require extension across different absorption edges, material classes, and synchrotron facilities. Systematic validation experiments coordinated through NFFA-Europe partners can identify where such techniques succeed and where they fail, informing algorithm improvements and guiding users on appropriate application domains.

Planned upgrades to the TriDAS architecture will replace the current ResNet50-based services with ViT architectures, leveraging NFFA-DI implementations already developed and validated. Additional enhancements will integrate Chapter 4 and 5 worfklows as operational notebooks within NFFA-DI.

Finally, extending technique coverage involves applying the FAIR-by-design methodology beyond the current scope of STM, SEM, and NEXAFS. Natural extensions include AFM, which exhibits similar curation challenges and artifact types; XPS, which faces comparable background removal requirements; Transmission Electron Microscopy (TEM), with distinct artifacts addressable through physics-informed restoration; and Raman spectroscopy, requiring analogous baseline removal and peak fitting.

7.3 Broader Impact and Closing Remarks

This thesis contributes to three interconnected domains influencing materials science research practices and infrastructure development. The FAIR-by-design methodology illustrated establishes practical templates for prospective data management in experimental laboratories, building upon FAIRification as an essential prerequisite phase. Common experimental settings require this two-phase progression: retrospective curation of existing archives followed by prospective embedding of FAIR principles in new workflows. The thesis' title "FAIR-by-Design Path" encompasses both phases, with FAIRification providing the necessary starting point for the design-oriented progression. Dissemination through NFFA-Europe partner institutions and FAIRmat consortium workshops ensures methodology transfer beyond originating laboratories.

The physics-informed ML paradigm addresses fundamental data scarcity challenges endemic to specialized experimental techniques. This methodology enables AI adoption in resource-constrained experimental domains where conventional supervised learning proves impractical, expanding ML utility from data-abundant to data-scarce scientific contexts.

Operational AI services deployed within European and Italian infrastructures demonstrate sustainable research-to-deployment pathways often absent in academic research emphasizing method publications over production implementations.

The integration of these themes establishes a template for workflow-focused AI in materials characterization, prioritizing practical experimental utility over benchmark performance metrics. Accessible services democratize advanced analysis capabilities across research institutions regardless of local computational resources or ML expertise. This workflow-centric perspective complements property prediction approaches dominating AI-materials science literature, addressing foundational challenges required before prediction tasks become feasible.

Future directions envision increasingly automated experimental workflows where AI systems handle routine data management, quality control, and analysis tasks while preserving human expertise for scientific interpretation, experimental design, and ambiguous case resolution. The detection-restoration integration planned exemplifies this human-AI collaboration: automated systems screen large archives and attempt computational remediation, flagging high uncertainty cases for expert review

rather than attempting fully autonomous decision-making. This balanced approach recognizes AI limitations while leveraging computational efficiency where appropriate, establishing sustainable workflows combining algorithmic capabilities with irreplaceable human domain knowledge. As materials characterization facilities generate ever-increasing data volumes, such integrated systems prove essential for extracting scientific value from experimental archives otherwise overwhelming manual analysis capacities.

Acknowledgments

I would like to thank my supervisors, Stefano Cozzini and Alberto Cazzaniga, for giving me the opportunity to undertake this project and for their constant encouragement and help. I also extend my gratitude to Alessio Ansuini, Matteo Biagetti and Luca Braglia for their invaluable contributions and collaboration.
I acknowledge the AREA Science Park supercomputing platform, ORFEO, made available for the research reported in this thesis, and the technical support of the staff of the Laboratory of Data Engineering.
This work was supported by the European Union – NextGenerationEU, M4C2, within the PNRR project NFFA-DI, CUP B53C22004310006, IR0000015, having benefited from the access provided by AREA Science Park in Trieste.

Bibliography

[1] J. Meyer, W. De Nolf, S. Debionne, S. Fisher, A. Götz, M. Guijarro, P. Guillou, A. Homs Puron, and V. Valls, "Facing the Challenges of Experiment Control and Data Management at ESRF-EBS," *JACoW*, vol. ICALEPCS2023, MO2AO01, 2023. DOI: 10.18429/JACoW-ICALEPCS2023-MO2AO01.

[2] R. Batra, L. Song, and R. Ramprasad, "Emerging materials intelligence ecosystems propelled by machine learning," *Nature Reviews Materials*, vol. 6, no. 8, pp. 655–678, 2021.

[3] M. D. Wilkinson, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, pp. 1–9, 2016. DOI: 10.1038/sdata.2016.18.

[4] P. Missier, K. Belhajjame, and J. Cheney, "The W3C PROV Family of Specifications for Modelling Provenance Metadata," 2013, pp. 773–776. DOI: 10.1145/2452376.2452478.

[5] T. Rodani, *Machine Learning techniques and visualization tools for STM images at CNR-IOM labs*, Sep. 2020. DOI: 10.5281/zenodo.5801169. [Online]. Available: https://doi.org/10.5281/zenodo.5801169.

[6] T. Rodani, E. Osmenaj, A. Cazzaniga, M. Panighel, A. Cristina, and S. Cozzini, "Towards the FAIRification of Scanning Tunneling Microscopy Images Open Access," *Data Intelligence*, vol. 5, no. 1, pp. 27–42, 2023.

[7] GO FAIR, *FAIRification Process*. [Online]. Available: https://www.go-fair.org/fair-principles/fairification-process/.

[8] S. Vigneri, "Design refinement and commissioning of a FAIR-by-design integrated data management system for an STM laboratory," *SISSA*, 2025.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 157–166.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, IEEE, 2009, pp. 248–255.

[13] A. Zaeemzadeh, N. Rahnavard, and M. Shah, "Norm-preservation: Why residual networks can become extremely deep?" *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3980–3990, 2020.

[14] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2021.

[15] M. H. Modarres, R. Aversa, S. Cozzini, R. Ciancio, A. Leto, and G. P. Brandino, "Neural Network for Nanoscience Scanning Electron Microscope Image Recognition," *Sci. Rep.*, pp. 1–12, 2017, ISSN: 2045-2322. DOI: 10.1038/s41598-017-13565-z. [Online]. Available: http://dx.doi.org/10.1038/s41598-017-13565-z.

[16] R. Aversa, P. Coronica, C. De Nobili, and S. Cozzini, "Deep Learning, Feature Learning, and Clustering Analysis for SEM Image Classification," *Data Intelligence*, pp. 513–528, 2020.

[17] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," in *European conference on computer vision*, Springer, 2016, pp. 241–257.

[18] T. Rodani, M. Panighel, C. Africh, and S. Cozzini, *Dataset of Scanning Tunneling Microscopy (STM) images of model surfaces for elementary steps in catalytic reactions*, version 1.0, Mar. 2024. DOI: 10.5281/zenodo.10886977. [Online]. Available: https://doi.org/10.5281/zenodo.10886977.

[19] T. Rodani, E. Osmenaj, A. Cazzaniga, M. Panighel, C. Africh, and S. Cozzini, *Dataset of Scanning Tunneling Microscopy (STM) images of graphene on nickel*, version 1.1, Dec. 2021. DOI: 10.5281/zenodo.7664070. [Online]. Available: https://doi.org/10.5281/zenodo.7664070.

[20] K. Belhajjame, R. B'Far, J. Cheney, S. Coppens, S. Cresswell, Y. Gil, P. Groth, G. Klyne, T. Lebo, J. McCusker, et al., "PROV-DM: The PROV Data Model," *W3C Recommendation*, vol. 14, pp. 15–16, 2013.

[21] P. Missier, K. Belhajjame, and J. Cheney, "The W3C PROV family of specifications for modelling provenance metadata," in *Proceedings of the 16th international conference on extending database technology*, 2013, pp. 773–776.

[22] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, "PROV-O: The PROV Ontology," *World Wide Web Consortium*, 2013.

[23] R. Aversa, A. Boubnov, D. De Angelis, C. Eschke, S. Irvine, R. E. Joseph, M. Kabbe, N. MacKinnon, M. Irene, M. Panighel, R. Thelen, and D. Valentinis, *The MDMC-NEP Glossary of Terms*, Feb. 2024. DOI: 10.5281/zenodo.10663833. [Online]. Available: https://doi.org/10.5281/zenodo.10663833.

[24] H. Trung Dong, *Prov Python*, 2014. [Online]. Available: https://github.com/trungdong/prov.

[25] European Organization For Nuclear Research and OpenAIRE, *Zenodo*, en, 2013. DOI: 10.25495/7GXK-RD71. [Online]. Available: https://www.zenodo.org/.

[26] T. Rodani, *t0m-R/STM_images 0.1.0*, version 0.1.0, Sep. 2020. DOI: 10.5281/zenodo.4019641. [Online]. Available: https://doi.org/10.5281/zenodo.4019641.

[27] A. Devaraju and R. Huber, *F-UJI - An Automated FAIR Data Assessment Tool*, version v1.0.0, Oct. 2020. DOI: 10.5281/zenodo.4063720. [Online]. Available: https://doi.org/10.5281/zenodo.4063720.

[28] M. Könnecke, F. A. Akeroyd, H. J. Bernstein, A. S. Brewster, S. I. Campbell, B. Clausen, S. Cottrell, J. U. Hoffmann, P. R. Jemian, D. Männicke, et al., "The NeXus data format," *Applied Crystallography*, vol. 48, no. 1, pp. 301–305, 2015.

[29] F. NeXus, *NXstm*, 2024. [Online]. Available: https://fairmat-nfdi.github.io/nexus_definitions/classes/contributed_definitions/NXstm.html#nxstm.

[30] W. Lo and J. Spence, "Investigation of STM image artifacts by in-situ reflection electron microscopy," *Ultramicroscopy*, vol. 48, no. 4, pp. 433–444, 1993. DOI: https://doi.org/10.1016/0304-3991(93)90119-I. [Online]. Available: https://www.sciencedirect.com/science/article/pii/030439919390119I.

[31] E. Van Loenen, D. Dijkkamp, A. Hoeven, J. Lenssinck, and J. Dieleman, "Evidence for tip imaging in scanning tunneling microscopy," *Applied physics letters*, vol. 56, no. 18, pp. 1755–1757, 1990.

[32] M. P. Yothers, A. E. Browder, and L. A. Bumm, "Real-space post-processing correction of thermal drift and piezoelectric actuator nonlinearities in scanning tunneling microscope images," *Review of Scientific Instruments*, vol. 88, no. 1, 2017.

[33] J. C. Straton, B. Moon, T. T. Bilyeu, and P. Moeck, "Removal of multiple-tip artifacts from scanning tunneling microscope images by crystallographic averaging," *Advanced Structural and Chemical Imaging*, vol. 1, no. 1, p. 14, 2015.

[34] T. Bilyeu, B. Moon, and P. Moeck, "Crystallographic Image Processing for Scanning Probe Microscopes," *Microscopy and Microanalysis*, vol. 18, no. S2, pp. 934–935, 2012.

[35] J. C. Straton, T. T. Bilyeu, B. Moon, and P. Moeck, "Double-tip effects on scanning tunneling microscopy imaging of 2D periodic objects: unambiguous detection and limits of their removal by crystallographic averaging in the spatial frequency domain," *Crystal Research and Technology*, vol. 49, no. 9, pp. 663–680, 2014.

[36] J. S. Villarrubia, "Algorithms for scanned probe microscope image simulation, surface reconstruction, and tip estimation," *Journal of research of the National Institute of Standards and Technology*, vol. 102, no. 4, p. 425, 1997.

[37] J. Welker and F. J. Giessibl, "Revealing the angular symmetry of chemical bonds by atomic force microscopy," *Science*, vol. 336, no. 6080, pp. 444–449, 2012.

[38] C. Chiutu, A. Sweetman, A. Lakin, A. Stannard, S. Jarvis, L. Kantorovich, J. Dunn, and P. Moriarty, "Precise orientation of a single C 60 molecule on the tip of a scanning probe microscope," *Physical review letters*, vol. 108, no. 26, p. 268 302, 2012.

[39] G. Schull, T. Frederiksen, A. Arnau, D. Sánchez-Portal, and R. Berndt, "Atomic-scale engineering of electrodes for single-molecule contacts," *Nature nanotechnology*, vol. 6, no. 1, pp. 23–27, 2011.

[40] D. Nečas and P. Klapetek, "Gwyddion: an open-source software for SPM data analysis," *Central European Journal of Physics*, vol. 10, no. 1, pp. 181–188, 2012.

[41] O. Gordon, P. D'Hondt, L. Knijff, S. Freeney, F. Junqueira, P. Moriarty, and I. Swart, "Scanning tunneling state recognition with multi-class neural network ensembles," *Review of Scientific Instruments*, vol. 90, no. 10, 2019.

[42] O. M. Gordon, F. L. Junqueira, and P. J. Moriarty, "Embedding human heuristics in machine-learning-enabled probe microscopy," *Machine Learning: Science and Technology*, vol. 1, no. 1, p. 015 001, 2020.

[43] A. Krull, P. Hirsch, C. Rother, A. Schiffrin, and C. Krull, "Artificial-intelligence-driven scanning probe microscopy," *Communications Physics*, vol. 3, no. 1, p. 54, 2020.

[44] I. Pitas, *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[46] N. Park and S. Kim, "How do vision transformers work?" *arXiv preprint arXiv:2202.06709*, 2022.

[47] N. L. Kolev, T. Rodani, N. J. Curson, T. J. Z. Stock, and A. Cazzaniga, *Generative Image Restoration and Super-Resolution using Physics-Informed Synthetic Data for Scanning Tunneling Microscopy*, 2025. arXiv: 2510.25921 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2510.25921.

[48] G. Binnig, H. Rohrer, C. Gerber, and E. Weibel, "7 × 7 Reconstruction on Si(111) Resolved in Real Space," *Phys. Rev. Lett.*, vol. 50, pp. 120–123, 2 Jan. 1983. DOI: 10.1103/PhysRevLett.50.120. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.50.120.

[49] M. Rost, L. Crama, P. Schakel, E. Van Tol, G. van Velzen-Williams, C. Overgauw, H. Ter Horst, H. Dekker, B. Okhuijsen, M. Seynen, et al., "Scanning probe microscopes go video rate and beyond," *Review of Scientific Instruments*, vol. 76, no. 5, 2005.

[50] L. Fricke, S. J. Hile, L. Kranz, Y. Chung, Y. He, P. Pakkiam, M. G. House, J. G. Keizer, and M. Y. Simmons, "Coherent control of a donor-molecule electron spin qubit in silicon," *Nature communications*, vol. 12, no. 1, p. 3323, 2021.

[51] B. Ramsauer, G. J. Simpson, J. J. Cartus, A. Jeindl, V. García-López, J. M. Tour, L. Grill, and O. T. Hofmann, "Autonomous single-molecule manipulation based on reinforcement learning," *The Journal of Physical Chemistry A*, vol. 127, no. 8, pp. 2041–2050, 2023.

[52] P. Leinen, M. Esders, K. T. Schütt, C. Wagner, K.-R. Müller, and F. S. Tautz, "Autonomous robotic nanofabrication with reinforcement learning," *Science advances*, vol. 6, no. 36, eabb6987, 2020.

[53] M. Rashidi and R. A. Wolkow, "Autonomous scanning probe microscopy in situ tip conditioning through machine learning," *ACS nano*, vol. 12, no. 6, pp. 5185–5189, 2018.

[54] D. S. Barker, P. J. Blowey, T. Brown, and A. Sweetman, "Automated Scanning Probe Tip State Classification without Machine Learning," *ACS nano*, vol. 18, no. 3, pp. 2384–2394, 2024.

[55] Z. Zhu, S. Yuan, Q. Yang, H. Jiang, F. Zheng, J. Lu, and Q. Sun, "Autonomous Scanning Tunneling Microscopy Imaging via Deep Learning," *Journal of the American Chemical Society*, vol. 146, no. 42, pp. 29 199–29 206, 2024.

[56] T. Rodani, A. Cazzaniga, et al., "Enhancing Multi-Tip Artifact Detection in STM Images Using Fourier Transform and Vision Transformers," in *ICML'24 Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024.

[57] F. Joucken, J. L. Davenport, Z. Ge, E. A. Quezada-Lopez, T. Taniguchi, K. Watanabe, J. Velasco Jr, J. Lagoute, and R. A. Kaindl, "Denoising scanning tunneling microscopy images of graphene with supervised machine learning," *Physical Review Materials*, vol. 6, no. 12, p. 123 802, 2022.

[58] J. Xie, W. Ko, R.-X. Zhang, and B. Yao, "Physics-augmented deep learning with adversarial domain adaptation: Applications to STM image denoising," *arXiv preprint arXiv:2409.05118*, 2024.

[59] M. M. Saad, R. O'Reilly, and M. H. Rehmani, "A survey on training challenges in generative adversarial networks for biomedical image analysis," *Artificial Intelligence Review*, vol. 57, no. 2, p. 19, 2024.

[60] K. Spruce, "Fabrication of Novel Atomic-Scale Electronic Devices from Dopants in Silicon," PhD thesis, University College London, London, UK, 2024.

[61] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[62] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[63] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[64] T. Rodani, N. Kolev, N. Curson, T. Stock, and A. Cazzaniga, *Dataset supporting: "Generative Image Restoration and Super-Resolution using Physics-Informed Synthetic Data for Scanning Tunneling Microscopy"*, Oct. 2025. DOI: 10.5281/zenodo.17474268. [Online]. Available: https://doi.org/10.5281/zenodo.17474268.

[65] Y. Liu, H. Zhou, B. Cui, W. Shang, and R. Lin, "Erase Diffusion: Empowering Object Removal Through Calibrating Diffusion Pathways," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 2418–2427.

[66] D. Fuoli, L. Van Gool, and R. Timofte, "Fourier space losses for efficient perceptual image super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2360–2369.

[67] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4641–4650.

[68] S. Shang, Z. Shan, G. Liu, L. Wang, X. Wang, Z. Zhang, and J. Zhang, "ResDiff: Combining CNN and Diffusion Model for Image Super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 8975–8983.

[69] R. C. Gonzalez, *Digital image processing*. Pearson Education India, 2009.

[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[71] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[72] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PmLR, 2021, pp. 8748–8763.

[73] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[74] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking fid: Towards a better evaluation metric for image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9307–9315.

[75] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[76] J. Croshaw, T. Dienel, T. Huff, and R. Wolkow, "Atomic defect classification of the H–Si (100) surface through multi-mode scanning probe microscopy," *Beilstein journal of nanotechnology*, vol. 11, no. 1, pp. 1346–1360, 2020.

[77] G. Panaccione, I. Vobornik, J. Fujii, D. Krizmancic, E. Annese, L. Giovanelli, F. Maccherozzi, F. Salvador, A. De Luisa, D. Benedetti, et al., "Advanced photoelectric effect experiment beamline at Elettra: A surface science laboratory coupled with Synchrotron Radiation," *Review of Scientific Instruments*, vol. 80, no. 4, 2009.

[78] C. Castán-Guerrero, D. Krizmancic, V. Bonanni, R. Edla, A. Deluisa, F. Salvador, G. Rossi, G. Panaccione, and P. Torelli, "A reaction cell for ambient pressure soft x-ray absorption spectroscopy," *Review of Scientific Instruments*, vol. 89, no. 5, 2018.

[79] L. Braglia, M. Fracchia, P. Ghigna, A. Minguzzi, D. Meroni, R. Edla, M. Vandichel, E. Ahlberg, G. Cerrato, and P. Torelli, "Understanding solid–gas reaction mechanisms by operando soft x-ray absorption spectroscopy at ambient pressure," *The Journal of Physical Chemistry C*, vol. 124, no. 26, pp. 14 202–14 212, 2020.

[80] L. Braglia, F. Tavani, S. Mauri, R. Edla, D. Krizmancic, A. Tofoni, V. Colombo, P. D'Angelo, and P. Torelli, "Catching the reversible formation and reactivity of surface defective sites in metal–organic frameworks: An operando ambient pressure-NEXAFS investigation," *The journal of physical chemistry letters*, vol. 12, no. 37, pp. 9182–9187, 2021.

[81] F. Tavani, M. Busato, D. Veclani, L. Braglia, S. Mauri, P. Torelli, and P. D'Angelo, "Investigating the high-temperature water/MgCl2 interface through ambient pressure soft X-ray absorption spectroscopy," *ACS Applied Materials & Interfaces*, vol. 15, no. 21, pp. 26 166–26 174, 2023.

[82] B. Ravel and M. Newville, "ATHENA, ARTEMIS, HEPHAESTUS: Data analysis for X-ray absorption spectroscopy using IFEFFIT," *Journal of Synchrotron Radiation*, vol. 12, no. 4, pp. 537–541, 2005. DOI: 10.1107/S0909049505012719.

[83] S. M. Webb, "SIXpack: a graphical user interface for XAS analysis using IFEFFIT," *Physica scripta*, vol. 2005, no. T115, p. 1011, 2005.

[84] E. Gann, C. R. McNeill, A. Tadich, B. C. Cowie, and L. Thomsen, "Quick AS NEXAFS Tool (QANT): a program for NEXAFS loading and analysis developed at the Australian Synchrotron," *Synchrotron Radiation*, vol. 23, no. 1, pp. 374–380, 2016.

[85] D. H. Simonne, A. Martini, M. Signorile, A. Piovano, L. Braglia, P. Torelli, E. Borfecchia, and G. Ricchiardi, "THORONDOR: a software for fast treatment and analysis of low-energy XAS data," *Synchrotron Radiation*, vol. 27, no. 6, pp. 1741–1752, 2020.

[86] J. J. Olivero and R. L. Longbothum, "Empirical fits to the Voigt line width: A brief review," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 17, no. 2, pp. 233–236, 1977. DOI: 10.1016/0022-4073(77)90161-3.

[87] B. T. Thole, P. Carra, F. Sette, and G. van der Laan, "X-ray circular dichroism as a probe of orbital magnetization," *Physical Review Letters*, vol. 68, no. 12, pp. 1943–1946, 1992. DOI: 10.1103/PhysRevLett.68.1943.

[88] F. De Groot and A. Kotani, *Core level spectroscopy of solids*. CRC press, 2008.

[89] J. Stöhr, *NEXAFS spectroscopy*. Springer Science & Business Media, 2013, vol. 25.

[90] R. Kurian, K. Kunnus, P. Wernet, S. M. Butorin, P. Glatzel, and F. M. de Groot, "Intrinsic deviations in fluorescence yield detected x-ray absorption spectroscopy: the case of the transition metal L2, 3 edges," *Journal of Physics: Condensed Matter*, vol. 24, no. 45, p. 452 201, 2012.

[91] C.-C. Chiu, Y.-W. Chang, Y.-C. Shao, Y.-C. Liu, J.-M. Lee, S.-W. Huang, W. Yang, J. Guo, F. M. de Groot, J.-C. Yang, et al., "Spectroscopic characterization of electronic structures of ultra-thin single crystal La0. 7Sr0. 3MnO3," *Scientific reports*, vol. 11, no. 1, p. 5250, 2021.

[92] T. Hu, Y. Xie, Y. L. Jin, and T. Liu, "A new method for extracting x-ray absorption fine structure and the atomic background from an x-ray absorption spectrum," *Journal of Physics: Condensed Matter*, vol. 9, no. 25, p. 5507, 1997.

[93] L. Tröger, D. Arvanitis, K. Baberschke, H. Michaelis, U. Grimm, and E. Zschech, "Full correction of the self-absorption in soft-fluorescence extended x-ray-absorption fine structure," *Physical Review B*, vol. 46, no. 6, pp. 3283–3289, 1992. DOI: 10.1103/PhysRevB.46.3283.

[94] A. Achkar, T. Regier, E. Monkman, K. Shen, and D. Hawthorn, "Determination of total x-ray absorption coefficient using non-resonant x-ray emission," *Scientific Reports*, vol. 1, no. 1, p. 182, 2011.

[95] A. Achkar, T. Regier, H. Wadati, Y.-J. Kim, H. Zhang, and D. Hawthorn, "Bulk sensitive x-ray absorption spectroscopy free of self-absorption effects," *Physical Review B—Condensed Matter and Materials Physics*, vol. 83, no. 8, p. 081 106, 2011.

[96] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[97] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018, pp. 334–340.

[98] J. Wahl, M. Sjödahl, and K. Ramser, "Single-step preprocessing of raman spectra using convolutional neural networks," *Applied spectroscopy*, vol. 74, no. 4, pp. 427–438, 2020.

[99] T. Chen, Y. Son, A. Park, and S.-J. Baek, "Baseline correction using a deep-learning model combining ResNet and UNet," *Analyst*, vol. 147, no. 19, pp. 4285–4292, 2022.

[100] M. Kazemzadeh, M. Martinez-Calderon, W. Xu, L. W. Chamley, C. L. Hisey, and N. G. Broderick, "Cascaded deep convolutional neural networks as improved methods of preprocessing raman spectroscopy data," *Analytical Chemistry*, vol. 94, no. 37, pp. 12 907–12 918, 2022.

[101] M. T. Gebrekidan, C. Knipfer, and A. S. Braeuer, "Refinement of spectra using a deep neural network: Fully automated removal of noise and background," *Journal of Raman Spectroscopy*, vol. 52, no. 3, pp. 723–736, 2021.

[102] D. Erb, *pybaselines: A Python library of algorithms for the baseline correction of experimental data*. DOI: 10.5281/zenodo.5608581. [Online]. Available: https://github.com/derb12/pybaselines.

[103] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman, "emcee: the MCMC hammer," *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 925, p. 306, 2013.

[104] NFFA-Europe, *NFFA-Europe: Nanoscience Foundries and Fine Analysis*, 2025. [Online]. Available: https://www.nffa.eu/.

[105] European Open Science Cloud (EOSC), *European Open Science Cloud (EOSC) – A Web of FAIR Data and Services for Science in Europe*, 2025. [Online]. Available: https://eosc.eu.

[106] NFFA-DI, *NFFA-DI: Nanoscience Foundries and Fine Analysis - Digital Infrastructure*, 2025. [Online]. Available: https://www.nffa-di.it/en/.

[107] European Strategy Forum on Research Infrastructures (ESFRI), *ESFRI Roadmap*, 2025. [Online]. Available: https://www.esfri.eu/esfri-roadmap.

[108] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay, et al., "Jupyter Notebooks-a publishing format for reproducible computational workflows.," in *ELPUB*, 2016, pp. 87–90.

[109] Area Science Park, *ORFEO: Open Research Facility for Epigenomics and Other*, 2025. [Online]. Available: https://orfeo-doc.areasciencepark.it/.

[110] A. B. Yoo, M. A. Jette, and M. Grondona, "Slurm: Simple linux utility for resource management," in *Workshop on job scheduling strategies for parallel processing*, Springer, 2003, pp. 44–60.

[111] S. A. Weil, S. A. Brandt, E. L. Miller, D. D. Long, and C. Maltzahn, "Ceph: A scalable, high-performance distributed file system," in *Proceedings of the 7th symposium on Operating systems design and implementation*, 2006, pp. 307–320.

[112] E. A. Brewer, "Kubernetes and the path to cloud native," in *Proceedings of the sixth ACM symposium on cloud computing*, 2015, pp. 167–167.

[113] M. Scheidgen, L. Himanen, A. N. Ladines, D. Sikter, M. Nakhaee, Á. Fekete, T. Chang, A. Golparvar, J. A. Márquez, S. Brockhauser, et al., "NOMAD: A distributed web-based platform for managing materials science research data," *Journal of Open Source Software*, vol. 8, no. 90, p. 5388, 2023.

[114] FAIRmat Consortium, *FAIRmat – FAIR Data Infrastructure for Condensed-Matter Physics and Chemical Physics of Solids*, 2025. [Online]. Available: https://www.fairmat-nfdi.eu/fairmat.

[115] S. A. Weil, A. W. Leung, S. A. Brandt, and C. Maltzahn, "Rados: A scalable, reliable storage service for petabyte-scale storage clusters," in *Proceedings of the 2nd international workshop on Petascale data storage: held in conjunction with Supercomputing'07*, 2007, pp. 35–44.

[116] M. B. Milligan, "Jupyter as common technology platform for interactive HPC services," in *Proceedings of the Practice and Experience on Advanced Research Computing: Seamless Creativity*, arXiv, 2018, pp. 1–6.

[117] Authentik Security Inc., *authentik: An open-source Identity Provider and Single Sign-On platform*, 2025. [Online]. Available: https://goauthentik.io.

[118] EasyDMP, *EasyDMP: A web application for Data Management Plans*, 2025. [Online]. Available: https://easydmp.areasciencepark.it.

[119] N. Carpi, A. Minges, and M. Piel, "eLabFTW: An open source laboratory notebook for research labs," *Journal of Open Source Software*, vol. 2, no. 12, p. 146, 2017. DOI: 10.21105/joss.00146. [Online]. Available: https://doi.org/10.21105/joss.00146.

[120] NFFA Europe - Trieste Advanced Data Services (TriDAS), *TriDAS - NFFA-Europe: Advanced Data Services for Nanoscale Research*, 2025. [Online]. Available: https://tridas.nffa.eu.

[121] The Kubernetes Ingress NGINX Community, *NGINX Ingress Controller for Kubernetes*, 2025. [Online]. Available: https://kubernetes.github.io/ingress-nginx/.

[122] R. Aversa, M. H. Modarres, S. Cozzini, R. Ciancio, and A. Chiusole, "Data descriptor: The first annotated set of scanning electron microscopy images for nanoscience," *Sci. Data*, vol. 5, 2018, ISSN: 20524463. DOI: 10.1038/sdata.2018.172.

[123] Keycloak Community, *Keycloak - Open Source Identity and Access Management*, 2025. [Online]. Available: https://www.keycloak.org.

[124] Helmholtz Metadata Collaboration. "Electron microscopy glossary." Accessed: 2025-12-18. [Online]. Available: https://emglossary.helmholtz-metadaten.de.

[125] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[126] A. Khalil and T. Rodani, *sem-meta: A unified Python package for SEM image processing, metadata extraction, OCR, and unit conversion*, 2025. [Online]. Available: `https://pypi.org/project/sem-meta/`.

[127] T. Rodani, *NFFA-DI Virtual Access Services*, 2025. [Online]. Available: `https://gitlab.com/area7/nffa-di/virtual-access-services`.

[128] T. Rodani, *Virtual Access ViT models*, 2025. [Online]. Available: `https://huggingface.co/t0m-R/`.

[129] Hugging Face Inc., *HuggingFace - The AI community building the future*, 2025. [Online]. Available: `https://huggingface.co`.