Data-Driven Methods for Accelerating Microstructure-Property Analysis in Metallic Materials

Von der Fakultät für Georessourcen und Materialtechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen

zur Erlangung des akademischen Grades einer

Doktorin der Ingenieurwissenschaften

genehmigte Dissertation

vorgelegt von

Chen Zhang, M.Sc.

Berichter: Univ.-Prof. Dr. Stefan Sandfeld
Univ.-Prof. Dr. rer. nat. Joachim Mayer

Tag der mündlichen Prüfung: 20.04.2026

Abstract

Understanding how microstructural features govern macroscopic properties is a central topic in materials science. Conventional experimental methods, although powerful, face limitations in resolving complex spatiotemporal phenomena and in providing sufficient data for quantitative modeling. This thesis addresses these challenges by applying materials data science approaches, including quantitative data mining of *in situ* TEM experiments and machine learning-based analysis of nanoindentation data, to enable rapid and quantitative materials characterization of metallic materials.

Two complementary research questions guided this work: (i) how to extract richer mechanistic insights and extend the interpretability of experimental observations through data-driven methodologies, and (ii) how to determine the sufficiency of experimental data for robust machine learning applications. To this end, two case studies are presented. The first uses *in situ* TEM to investigate dislocation dynamics in a Cantor alloy. Treating dislocations as probes of the local energy landscape and applying data-mining techniques allows for the quantitative characterization of the strength, evolution, and spatial distribution of pinning points. The results show that pinning strength varies as dislocations pass, that pinning sites can shift position, and that their spatial distribution is heterogeneous. These findings help explain how microstructural obstacles influence plastic deformation and hardening. The second study applies machine learning to analyze nanoindentation data from Cu–Cr composites with controlled heterogeneity. A Gaussian mixture model is used to identify mechanical phases and estimate their volume fractions directly from experimental data. More importantly, a cross-validation framework is introduced to assess how much data is required for stable model performance, providing practical guidance for experimental data collection.

Together, these studies illustrate how data-driven analysis can make experimental data more efficient to use and more informative in revealing underlying mechanisms. The methods developed in this thesis strengthen the quantitative link between microstructure and properties and, in doing so, support the ongoing transition of materials science toward a data-driven paradigm with practical routes for accelerated discovery and design. The main results of Chapters 6 and 7 have been published as follows:

- C. Zhang, H. Song, D. Oliveros, A. Fraczkiewicz, M. Legros, and S. Sandfeld, *Data-mining of in-situ TEM experiments: On the dynamics of dislocations in CoCrFeMnNi alloys*, Acta Materialia 241 (2022), 118394.
- C. Zhang, C. Bos, S. Sandfeld, and R. Schwaiger, *Unsupervised learning of nanoindentation data to infer microstructural details of complex materials*, Frontiers in Materials 11 (2024), 1440608.

Überblick

Das Verständnis, wie Mikrostrukturen makroskopische Eigenschaften bestimmen, ist ein Kernthema der Materialwissenschaften. Konventionelle Methoden sind zwar leistungsfähig, stoßen jedoch an Grenzen bei der Auflösung komplexer raumzeitlicher Phänomene und bei der Bereitstellung ausreichender Daten für quantitatives Modellieren. Diese Dissertation begegnet diesen Herausforderungen mit datenwissenschaftlichen Ansätzen, darunter die quantitative Analyse von *in situ* TEM Experimenten und maschinelles Lernen für Nanoindentationsdaten, um eine schnelle und präzise mechanische Charakterisierung metallischer Werkstoffe zu ermöglichen.

Zwei Fragen leiteten die Arbeit: (i) wie sich reichhaltigere mechanistische Einsichten gewinnen und die Interpretierbarkeit experimenteller Beobachtungen durch datengetriebene Methoden erweitern lassen, und (ii) wie viel Daten für robuste Anwendungen des maschinellen Lernens erforderlich sind. Dazu werden zwei Fallstudien vorgestellt. Die erste untersucht mit *in situ* TEM die Versetzungsdynamik in einer Cantor Legierung. Versetzungen dienen als Sonden der lokalen Energielandschaft; Data-Mining ermöglicht die quantitative Charakterisierung von Stärke, Entwicklung und Verteilung von Pinning-Zentren. Die Ergebnisse zeigen variable Pinning-Stärken, verschiebbare Positionen und eine heterogene Verteilung—Befunde, die erklären, wie mikrostrukturelle Hindernisse plastische Verformung und Verfestigung beeinflussen. Die zweite Studie nutzt maschinelles Lernen zur Analyse von Nanoindentationsdaten aus Cu–Cr Verbundwerkstoffen. Ein Gaußsches Mischungsmodell identifiziert mechanische Phasen und schätzt ihre Volumenanteile, während ein Cross-Validation Framework den nötigen Datenumfang für stabile Modellleistungen bewertet und Leitlinien für die Datenerhebung liefert.

Zusammen zeigen diese Studien, wie datengetriebene Analysen experimentelle Daten effizienter nutzbar und informativer für das Verständnis zugrunde liegender Mechanismen machen. Die entwickelten Methoden stärken die quantitative Verknüpfung zwischen Mikrostruktur und Eigenschaften und fördern den Wandel der Materialwissenschaften hin zu einem datengetriebenen Paradigma mit praxisnahen Ansätzen für beschleunigte Materialentdeckung und -entwicklung. Die Hauptergebnisse der Kapitel 6 und 7 wurden wie folgt veröffentlicht:

- C. Zhang, H. Song, D. Oliveros, A. Fraczkiewicz, M. Legros, and S. Sandfeld, *Data-mining of in-situ TEM experiments: On the dynamics of dislocations in CoCrFeMnNi alloys*, Acta Materialia 241 (2022), 118394.
- C. Zhang, C. Bos, S. Sandfeld, and R. Schwaiger, *Unsupervised learning of nanoindentation data to infer microstructural details of complex materials*, Frontiers in Materials 11 (2024), 1440608.

Acknowledgments

This thesis was made possible by generous support throughout my doctoral studies, with the research funded by the European Research Council through the ERC Grant Agreement No. 759419 MuDiLingo ("A Multiscale Dislocation Language for Data-Driven Materials Science") and further supported by CEMES-CNRS through access to microscopy facilities. I am deeply grateful for this institutional and financial support.

First, I am sincerely grateful to my supervisor, Prof. Dr. Stefan Sandfeld, for his guidance, encouragement, and confidence throughout this project. His high standards and scientific insight have constantly inspired me to strive for improvement, while his generous support and trust in my abilities have enabled me to develop as an independent researcher.

With special thanks to Dr. Marc Legros, Dr. Hengxu Song, and Prof. Dr. Ruth Schwaiger, I would also like to acknowledge all our collaborators for their constructive discussions, insightful comments, and valuable contributions to the models. Their input has had a lasting impact on this thesis.

I am equally indebted to my colleagues from the Institute of Materials Data Science and Informatics (IAS-9), FZ Jülich, especially Dr. Ing. Ahmad Zainul Ihsan, Dr. Ing. Kishan Govind, and many others for their stimulating conversations, collaborative spirit, and the pleasant atmosphere that made the research process rewarding and enjoyable.

Finally, my heartfelt thanks go to my family and friends. Their patience, encouragement, and unwavering support gave me the energy to continue when challenges arose. This work stands as a tribute to all who have supported me along this journey.

Acronyms and abbreviations

The following acronyms and abbreviations are used:

Acronyms (general)

TEM	transmission electron microscopy	4
ML	machine learning	5
AI	artificial intelligence	5
MGI	Materials Genome Initiative	5
AFLOW	Automatic Flow	5
OQMD	Open Quantum Materials Database	5
NOMAD	Novel Materials Discovery	5
MCCA	multi-component complex alloys	17
fcc	face-centered-cubic	17
hcp	hexagonal close-packed	18
STEM-EDS	scanning transmission electron microscopy with energy-dispersive X-ray spectroscopy	18
APT	atom probe tomography	18
HEA	high entropy alloy	19
ISE	indentation size effect	23
CNN	convolutional neural network	23
EM	expectation-maximization	23
GMM	Gaussian mixture model	23
PSP	processing-structure-property	25
OpenKIM	open knowledgebase of interatomic models	26
AMANDA	autonomous materials and device application platform	28
RD	rolling direction	31
TD	transverse direction	31
ND	normal direction	31
EELS	electron energy loss spectroscopy	45

CDD continuum dislocation dynamics 47
DDD discrete dislocation dynamics 47
BIC Bayesian information criterion 61
PCA principal component analysis 62
ARI adjusted Rand index 63
FIB focused ion beam 84
PDF probability density function 97

Abbreviations (project-specific)

SCS sample coordinate system 31
CCS crystal coordinate system 31
WCS world coordinate system 32
CG coarse graining 59
D density 68
E modulus of elasticity 68
TS tensile strength 68
TC thermal conductivity 68
CTE coefficient of linear expansion 68
ER electrical resistivity 68
Eab elongation at break 68

Nomenclature

$\boldsymbol{b}$	Burgers vector
$\boldsymbol{s}$	Slip direction
$\boldsymbol{n}$	Slip plane normal vector
T	Line tension force
r_0	Dislocation core radius
R_{ext}	Outer cutoff radius (distance from dislocation core)
R_{c}	Radius of curvature of a bent dislocation line
r_0	Core radius of a dislocation
E	Young's modulus (elastic modulus)
H	Hardness
E_{el}	Elastic energy of a dislocation
ν	Poisson's ratio
$\boldsymbol{f}$	Force vector
$\boldsymbol{f}_{\text{obs}}$	Obstacle (pinning) force vector
α	Angle between Burgers vector and dislocation line
α_1, α_2	Angles of dislocation segments with respect to horizontal
$\boldsymbol{\tau}$	Shear stress
k	Curvature (geometry)
α_j	Weight of the j-th Gaussian component
$\boldsymbol{\theta}_j$	Parameter vector of the j-th Gaussian component
$\boldsymbol{\mu}_j$	Mean vector of the j-th Gaussian component
$\boldsymbol{\Sigma}_j$	Covariance matrix of the j-th Gaussian component
$\boldsymbol{\Phi}_k$	Set of parameters of a mixture model with k components
$\boldsymbol{x}_i$	i-th feature vector in the dataset
d	Dimensionality of the feature space
$F(\boldsymbol{x} \mid \boldsymbol{\Phi}_k)$	Mixture model probability density function
y_i	Cluster label assigned to data point i
$\boldsymbol{Y}$	Set (vector) of all predicted cluster labels
A_k	Clustering algorithm producing k clusters
$\mathcal{D}$	Dataset
$\mathcal{D}_{\text{train}}$	Training dataset
$\mathcal{D}_{\text{test}}$	Testing dataset
N	Number of data records
k	Number of mixture components (in GMM) or number of folds (CV)
Δ	Difference operator; Laplacian
Ω	Hypothesis space; domain or region

Conventions

Throughout this thesis, we use the following conventions:

- **Scalars**: Scalars (ordinary numbers or variables) are written in *italic* Latin or Greek letters, e.g. t, m, E, α, λ.
- **Vectors**: Vectors are denoted by bold *italic* lowercase letters, e.g. $\boldsymbol{v}, \boldsymbol{e}$; Greek vectors are written in bold as well, e.g. $\boldsymbol{\sigma}$.
- **Matrices and Tensors**:
  - Matrices are denoted by bold uppercase sans-serif letters, e.g. $\mathbf{A}, \mathbf{X}$.
  - Higher-order tensors follow the same convention, written as bold uppercase sans-serif letters with indices, e.g. $\mathbf{T}_{ijk}$.
- **Sets and Collections**: The sets of real and natural numbers are denoted by $\mathbb{R}, \mathbb{N}$, respectively. Datasets, partitions, and families of distributions are denoted in calligraphic script, e.g. $\mathcal{D}$ (dataset), $\mathcal{C}$ (class labels), $\mathcal{K}$ (cluster assignment).
- **Functions**: Functions are written as $f(x)$, where f denotes the function and x the variable. For example, $f(x) = x^2 + 1$. If fixed parameters are required, we may write $f(x; a, b) = ax + b$.
- **Brackets**:
  - Round brackets ( ): grouping and function arguments.
  - Square brackets [ ]: commutators or operator definitions, e.g. $[x, p] = i\hbar$.
  - Curly brackets { }: sets, or anticommutators, e.g. $\{a, b\}$.
  - Angular brackets ⟨ ⟩: an averaged quantity over the ensemble, e.g. $\langle \boldsymbol{v} \rangle(\boldsymbol{r})$.
- **Superscripts and Subscripts**: Used for exponents, indices, or special notation, e.g. $\sigma^2, \mu_1, \boldsymbol{x}_{\mathrm{w}}$.
- **Greek Letters**: Frequently used for variables and constants, e.g. α, Δ, Ω. In applied contexts such as machine learning, certain Greek letters carry fixed meanings due to widely accepted conventions.
- **Special Conventions**: Rotations in 3D space are parameterized by three Euler angles (ϕ_1, Φ, ϕ_2), representing successive intrinsic rotations about the z-, x-, and z-axes, respectively.

Contents

List of Figures

Part I.

INTRODUCTION AND FOUNDATIONS

1. Introduction

1.1. The Integration of Materials Science and Data Science

The fundamental objective of materials science is to unravel the intrinsic relationships between microstructural features and macroscopic properties of materials [1]. Microstructural characterization plays a crucial role in this pursuit, providing the scientific foundation for interpreting and quantifying these relationships. Traditional experimental approaches, such as electron microscopy and spectroscopic analysis, have significantly advanced our ability to probe material structures. However, they often remain limited in capturing complex spatiotemporal phenomena such as internal stress fields, velocity fields, or energy distributions [2].

The development of scientific methodology has undergone major paradigm shifts over the centuries [3]. Figure 1.1 presents the four foundational paradigms of science, empirical, theoretical, computational, and data-driven, each marking a milestone in how knowledge is generated and applied. In this thesis, these paradigms are not treated as isolated phases but as a cumulative and interdisciplinary foundation for modern materials research.

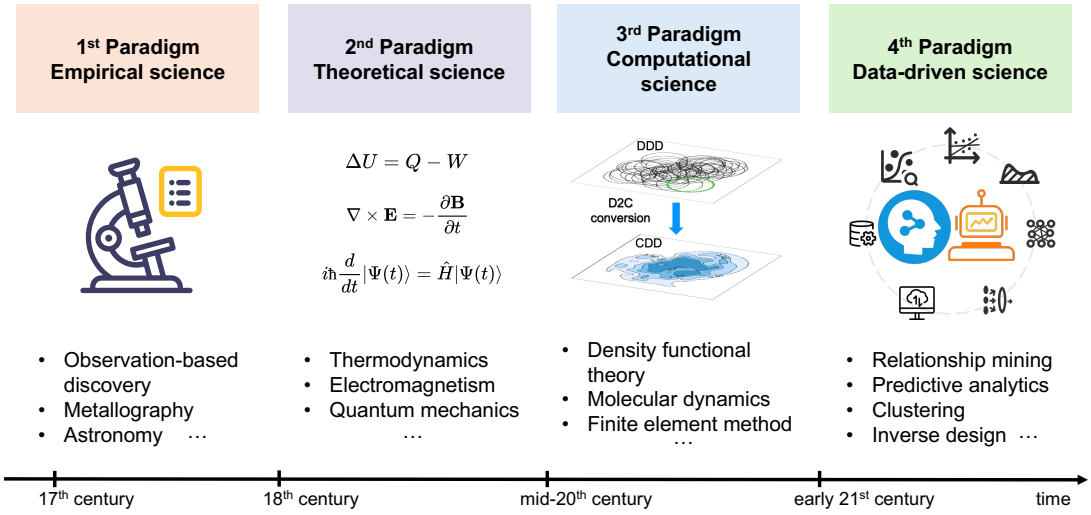

Figure 1.1. The four paradigms of science and their historical evolution: empirical (observation), theoretical (analytical laws), computational (modeling, images adapted from Sandfeld et al. [4]), and data-driven (pattern extraction from large datasets).

Emerging in the 17th century, the first paradigm laid the groundwork for modern scientific methodology [5]. Galileo Galilei, for example, shifted natural philosophy from speculative reasoning to systematic experimentation by quantifying motion through inclined-plane experiments [6]. In materials science, empirical observations drove the early development of phase diagrams, metallography, and crystallography [7]. In this thesis, techniques like transmission electron microscopy (TEM) [8] and nanoindentation [9] continue to provide high-resolution empirical data, which serve as the foundation for data-driven analysis.

In the 18th and 19th centuries, physics matured into a discipline governed by analytical laws. Thermodynamics, electromagnetism, and quantum mechanics allowed researchers to model physical phenomena with increasing precision [10, 11]. In materials science, this enabled predictive descriptions of defect behavior, stress-strain responses, and phase transformations. Theoretical concepts such as the dislocation, Burgers vector or energy barriers [12, 13], serve as the foundation for the interpretation of experimental results in this study.

By the mid-20th century, the advent of computational capabilities ushered in a new paradigm: computational science [14, 15]. As scientific theories grew increasingly complex, the limitations of purely analytical models became apparent, and computational approaches emerged as a powerful alternative. Techniques such as density functional theory (DFT) [16], molecular dynamics (MD) [17], and finite element modeling (FEM) [18] revolutionized materials research by enabling multiscale simulations. In this thesis, computational methods are employed to connect discrete dislocation configurations with continuum-level descriptions via dislocation geometries, providing a clear example of the fusion of geometry, numerical computation, and physics.

The early 21st century witnessed an unprecedented proliferation of digital data and internet connectivity, creating new opportunities for data-driven analysis across diverse scientific disciplines. Concurrently, materials science and engineering began to accumulate vast datasets from computational simulations, experiments, and theoretical studies. Together, these developments enabled the exploration of microstructure-property relationships from a broader data perspective, thereby fostering the design of novel materials [19]. This shift gave rise to materials data-driven science, which emphasizes extracting patterns, correlations, and predictive rules directly from data—often without explicit physical models [20]. In this context, the present thesis implements this paradigm through data mining [21, 22], in particular unsupervised learning, yielding new insights where direct measurements or analytical models are inadequate, while maintaining a consistent link to the physical interpretation of the underlying processes.

This thesis integrates all four paradigms: empirical measurements, theoretical modeling, computational implementation, and data-driven analysis. Together, they form the methodological foundation for a hybrid, data-driven framework for materials characterization developed in this work.

1.2. The Rise of Data-Driven Approaches in Materials Science

Since the early 21st century, the materials science community has witnessed the growing adoption of data-driven techniques to overcome the limitations of traditional approaches. Major initiatives, most notably the Materials Genome Initiative (MGI) [23], have fostered the development of extensive materials databases, such as the Materials Project [24], which in turn accelerated the integration of computation, data, and materials discovery. This database encompasses information on over 100,000 materials, including crystal structures, electronic properties, and mechanical attributes. Subsequently, similar databases and platforms, such as the Automatic Flow (AFLOW) [25], the Open Quantum Materials Database (OQMD) [26], and the Novel Materials Discovery (NOMAD) [27], were developed, vastly expanding the data resources available to materials scientists. Together, these developments mark a paradigm shift from case-by-case studies toward data-intensive discovery in materials science.

The emergence of Industry 4.0 concepts [28], coupled with advances in high-throughput computing, has further accelerated the integration of machine learning (ML) and artificial intelligence into materials science workflows [29, 30]. Techniques such as high-throughput experiments, multiscale modeling, and automated data analysis have significantly accelerated materials research, enabling rapid screening of large material datasets, prediction of material properties, and simulation of materials behavior across various length and time scales [31].

While these initiatives and technological advances have significantly accelerated progress, they also expose fundamental challenges that must be addressed for data-driven methods to reach their full potential.

1.3. Challenges and Research Questions in Materials Data Science

The rapid growth of data-driven approaches in materials science has created significant opportunities but also introduced critical challenges [32]. Experimental and computational studies now produce vast, heterogeneous datasets that often exceed the capacity of conventional analytical methods. This section highlights the key challenges commonly encountered in data-driven materials science and specifies those that are directly addressed in this thesis:

- **Extracting complex microstructural features from experimental data:** Many key properties, such as internal stresses or velocity distributions, remain difficult to capture directly with existing experimental techniques.
- **Standardizing and integrating data across domains:** Variability in experimental conditions, measurement protocols, and computational models complicates data integration and hinders reproducibility.

- **Improving interpretability of machine-learning models:** Although ML methods can predict material performance, their black-box nature often limits the extraction of mechanistic insights.
- **Balancing data quantity and model reliability:** Determining how much experimental data is sufficient for robust ML predictions remains challenging, particularly in mechanical testing where datasets are typically scarce.

These challenges motivate the central research questions addressed in this thesis:

(i) **How can data-driven methods overcome the spatiotemporal limitations of conventional experimental techniques to enable more comprehensive microstructural characterization?** This question directly responds to the challenge that many key properties, such as internal stresses or velocity distributions, remain inaccessible to current experimental techniques. Addressing it requires integrating experimental data with computational modeling and data-mining techniques to access quantities beyond the reach of direct measurement.

(ii) **How can data-driven methods be employed to uncover and quantify the fundamental relationships between microstructure and mechanical properties?** This question highlights the need to move beyond black-box clustering by improving interpretability and, for example, linking unsupervised learning results to materials understanding. It also involves determining how much and what type of data are sufficient for robust models, with the goal of reducing experimental effort and time while still capturing the essential structure-property relationships.

These questions call for a cross-disciplinary methodology combining experimental data, computational modeling, and machine learning—a framework that is implemented and tested in the remainder of this thesis. The next section translates them into the specific objectives and scope of this study.

1.4. Research Objectives and Scope of the Thesis

To address these challenges and answer the resulting research questions, this research aims to develop a data-driven microstructural characterization framework that enhances traditional experimental techniques. This thesis presents two case studies:

(I) Data-Driven Characterization of Dislocation Behavior in Cantor Alloys

Using TEM, this study examines the energy field variations associated with dislocation motion. Dislocation microstructures serve as probes to investigate pinning points and their impact on dislocation dynamics. This analysis underscores the significance of data management and standardization in experimental

research, facilitating the quantification of pinning point strength variations. By leveraging data-driven techniques, the study establishes structure-property correlations that inform material optimization and design.

(II) Optimal Data Utilization for Machine Learning in Microstructure-Property Analysis

This study explores the critical question of determining the optimal data quantity necessary for effectively applying machine learning to microstructure-property analysis. Through nanoindentation experiments, it investigates the size effect in copper-chromium (Cu–Cr) composites by analyzing mechanical phase variations at different depths.

By combining experimental data with data-driven methodologies, this thesis addresses two complementary challenges, namely: (i) extending experimental observations to phenomena that are otherwise beyond direct measurement, and (ii) assessing the sufficiency of input data for reliable machine-learning applications in microstructure-property analysis. The resulting framework enhances the scope and efficiency of microstructural characterization, thereby enabling a more reliable assessment of structure–property relationships. Such advances ultimately pave the way for developing robust strategies in materials design and optimization.

1.5. Scope and Outline of the Thesis

This thesis focuses on advancing data-driven microstructural characterization, with particular emphasis on dislocation behavior and microstructure-property relationships. The scope of the work is defined by the integration of experimental input, computational modeling, and machine learning methods to address the research challenges identified in Section 1.3. Accordingly, the thesis is organized to move from theoretical foundations, through methodological developments, to application-oriented studies:

Chapter 2 provides a comprehensive review of theoretical frameworks, including dislocation theories, nanoindentation principles, and materials data science. This establishes the scientific background necessary for the methodological and application-focused parts of the thesis.

Chapter 3 introduces a 3D reconstruction technique for dislocation analysis using TEM, forming the basis for quantifying dislocation structures from experimental data.

Chapter 4 presents a data-driven framework for computational characterization of dislocation dynamics, bridging discrete and continuum-level descriptions and laying the groundwork for later case studies.

Chapter 5 develops a probabilistic clustering approach, centered on Gaussian mixture models, to analyze microstructural datasets. This chapter also introduces evaluation strategies for unsupervised learning models.

Chapter 6 applies the developed framework to *in situ* TEM experiments on Cantor alloys, extracting quantitative information on dislocation pinning points and their spatio-temporal evolution.

Chapter 7 employs Gaussian mixture models to analyze nanoindentation data from Cu–Cr composites, highlighting how unsupervised learning can uncover size effects and microstructural heterogeneity.

Chapter 8 summarizes the key contributions of the thesis, reflects on the integration of data-driven methods with experimental and computational materials science, and outlines perspectives for future research.

Taken together, this structure ensures a coherent progression from fundamentals to applications, systematically addressing the research objectives posed in Chapter 1.

2. Theoretical Foundations

This chapter outlines the theoretical principles that form the foundation of the data-driven approaches explored in this work. It begins with dislocation-based crystal plasticity, which describes how metals accommodate plastic deformation through dislocations, thereby linking microstructural mechanisms to macroscopic strength.

The discussion then focuses on nanoindentation as a small-scale mechanical testing technique. Built upon classical contact mechanics, hardness theory, and the Oliver–Pharr analysis, nanoindentation enables the quantitative evaluation of elastic and plastic properties directly from load–displacement responses.

Finally, the chapter introduces key concepts in materials data science. It discusses the goals of data-driven materials research, the infrastructure that enables such approaches, and recent progress in applying such approaches to materials characterization, highlighting how data-driven methods accelerate materials characterization and advance the understanding of structure–property relationships.

2.1. Dislocation-based Crystal Plasticity

2.1.1. Description of a Dislocation

The concept of dislocations in crystals began to take form in the early twentieth century. In the 1920s, Frenkel [33] proposed a theoretical model for slip in a single perfect crystal, arguing that the critical shear stress required for slip is governed by interatomic forces between atomic planes. However, his estimate of the required shear stress exceeded experimental values by an order of 10^3–10^4, indicating that a key physical mechanism was missing from the theory.

In 1934, three scientists, G. I. Taylor [34], E. Orowan [35], and M. Polanyi [36], independently and simultaneously proposed the existence of defects in crystals (in contrast to the earlier assumption of perfect crystals) and demonstrated that plastic flow or deformation could occur via the motion of dislocations, rather than the collective motion of entire atomic planes. This theory successfully explained the large discrepancy between the theoretical magnitude of shear strength and experimental observations, and it was also the first to distinguish two fundamental forms of plastic deformation: edge dislocations and screw dislocations, the latter proposed by Burgers in 1939 [37].

Building upon these foundational ideas, the two primary types of dislocations can be described in terms of their distinct atomic configurations and the resulting lattice distortions (see Figure 2.1). Edge dislocations can be visualized as being caused by

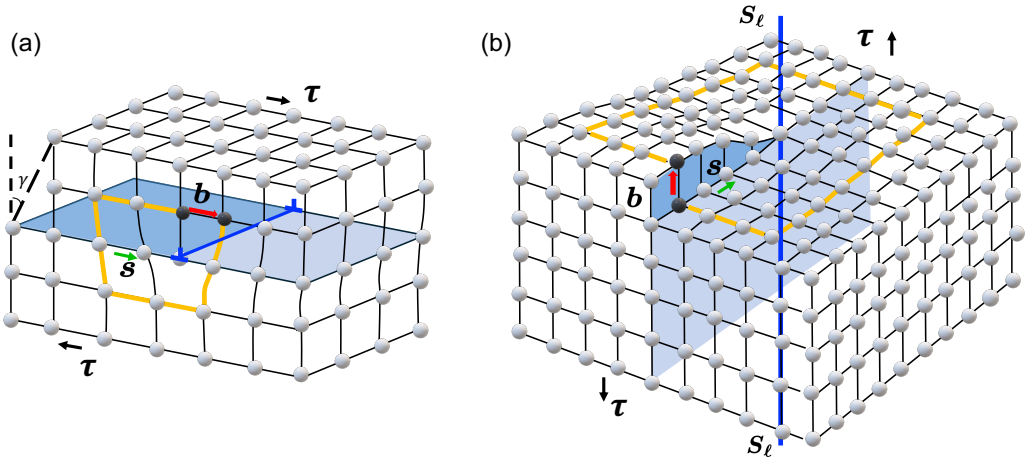

Figure 2.1. Plastic deformation under shear stress. (a) Intermediate stage of slip of a positive edge dislocation. (b) Intermediate stage of slip of a right-handed screw dislocation. Blue lines indicate the dislocation, black arrows indicate shear stress $\boldsymbol{\tau}$, green arrows indicate slip direction $\boldsymbol{s}$, and red arrows indicate the Burgers vector $\boldsymbol{b}$. The yellow lines represent the Burgers circuit. When the dislocation leaves the crystal, the crystal has undergone a shear strain γ.

the termination of a plane of atoms in the middle of a crystal. The dislocation line (in Figure 2.1a) separates the slipped (darker blue area) from the unslipped portion (lighter blue area) of the crystal. On either side of the dislocation line, the atoms may be slightly displaced, causing a local distortion of the lattice. Screw dislocations, on the other hand, are crystallographic line defects in which the crystal planes on either side of the dislocation are displaced relative to each other by a shear parallel to the dislocation line. As a result, the atomic planes wrap around the dislocation core in a helical arrangement (in Figure 2.1b), giving rise to a spiral mismatch in the lattice. While edge and screw dislocations represent the two idealized limiting cases, most real dislocations fall somewhere in between. Dislocations whose line directions are neither parallel nor perpendicular to the Burgers vector are referred to as mixed dislocations, exhibiting both edge and screw character.

To describe the magnitude and direction of the lattice distortion associated with a dislocation, the concept of the Burgers vector $\boldsymbol{b}$ was introduced by the Dutch scientist Jan Burgers in the 1930s [38]. As depicted in Figure 2.1, $\boldsymbol{b}$ is a lattice translation vector connecting two points within the crystal lattice. It is determined using a Burgers circuit (highlighted in yellow), which involves traversing a closed loop around a dislocation and comparing it to the same loop in a perfect crystal. A dislocation that possesses a Burgers vector aligned with a lattice translation vector is commonly referred to as a perfect or unit dislocation. The Burgers vector is always perpendicular to the dislocation line for edge dislocations and parallel to it for screw dislocations. In reality, the dislocation line is often oriented at an arbitrary angle to $\boldsymbol{b}$, exhibiting both edge and screw character.

While $\boldsymbol{b}$ provides a geometric description of a single dislocation, the collective

dynamics of many dislocations are crucial for understanding how microscopic lattice defects give rise to macroscopic deformation. As shown in Figure 2.2, even a small number of dislocations gliding through the lattice under tensile loading can introduce steps on the surface, slightly altering the specimen's dimensions. Although the displacement associated with the passage of a single dislocation is extremely small ($\sim 10^{-10}$ m), the collective motion of a large number of dislocations can lead to significant shape changes in a crystal. From a mechanistic perspective, plastic deformation originates from shear processes facilitated by the motion of dislocations. Slip occurs preferentially along specific crystallographic planes (known as slip planes), which present lower resistance to dislocation motion compared with other planes in the lattice. The selection of active slip planes is governed by the crystal structure, with each structure exhibiting preferred planes for dislocation motion. The orientation of these planes, combined with the Burgers vector $\boldsymbol{b}$, determines the direction and magnitude of plastic deformation.

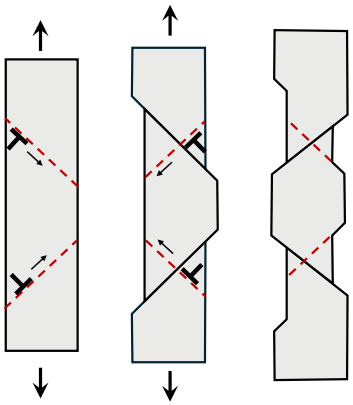

Figure 2.2. Geometrical representation of slip occurring within a cylindrical crystal. As the average atomic spacing is unaffected, the total volume does not change (adapted from Ashby et al. [39]).

By the mid-20th century, dislocation theory had become a central framework for explaining plastic deformation. Yet, the existence of dislocations was supported mainly by indirect evidence and theoretical inference [40, 41, 42, 43, 44] until the mid-1950s, when the advent of transmission electron microscopy (TEM) enabled their direct imaging in crystal lattices [45, 46]. By transmitting a focused electron beam through an ultrathin specimen, TEM resolves internal features such as dislocation lines at the nanoscale. This advancement provided crucial experimental validation for earlier theoretical models, transforming dislocation theory from a speculative concept into a directly observable reality. Prior to this, theories of crystal strength relied largely on continuum mechanics and empirical relations such as the Hall–Petch law [47, 48], which lacked a clear link between microstructure and mechanical behavior. The emergence and subsequent confirmation of dislocation theory filled this gap by offering a physics-based explanation for key phenomena such as yield strength, work hardening,

creep, and fracture [49, 50, 12]. Today, with advances in imaging and computation, dislocations are recognized not only as carriers of plastic deformation but also as essential tools for probing the mechanical behavior of materials.

2.1.2. The Stress Field around a Dislocation

While the geometric and crystallographic descriptions presented above capture the essential nature of dislocations and their role in plastic deformation, it is important to recognize that dislocations are not merely geometric defects. Their presence fundamentally alters the local stress state of the crystal. Dislocations in a crystal lattice create stress fields around them due to the associated distortion and rearrangement of atoms. Quantifying the elastic stress field surrounding dislocations presents a challenging task. For simplicity, the elastic continuum model is typically employed, whereby the crystal containing dislocations is treated as a homogeneous, continuous, and isotropic elastic medium.

Various theories, such as Volterra's classical treatment [49], can be used to characterize the stress fields generated by dislocations. These fields arise from the lattice distortion associated with dislocations and act to resist further deformation, tending to restore the material toward its original state. The continuum solution provides a good approximation of the stress and strain fields at distances sufficiently far from the dislocation core. Despite its limitations, this approach remains highly valuable for understanding dislocation interactions and their influence on mechanical behavior.

The mechanics of materials establishes that the stress state within a solid can be described by nine components of the stress tensor $\boldsymbol{\sigma}$ at any point. Consider an isotropic hollow cylinder cut along the xz-plane. If the two surfaces are displaced by a Burgers vector $\boldsymbol{b}$ relative to each other along the z-axis and rejoined, a screw dislocation is formed (Figure 2.3a).

The stress field of a straight screw dislocation in an infinite medium is given by

$$\begin{aligned} \sigma_{xx} &= \sigma_{yy} = \sigma_{zz} = \sigma_{xy} = \sigma_{yx} = 0 \\ \sigma_{xz} &= \sigma_{zx} = -\frac{Gb}{2\pi}\frac{y}{x^2+y^2} \\ \sigma_{yz} &= \sigma_{zy} = \frac{Gb}{2\pi}\frac{x}{x^2+y^2} \end{aligned} \tag{2.1}$$

where $b = \|\boldsymbol{b}\|$. Similarly, if the cut is made such that the two sides are displaced by $\boldsymbol{b}$ along the x-axis, an edge dislocation is formed (Figure 2.3b). The stress field of a straight edge dislocation in an infinite medium is expressed as

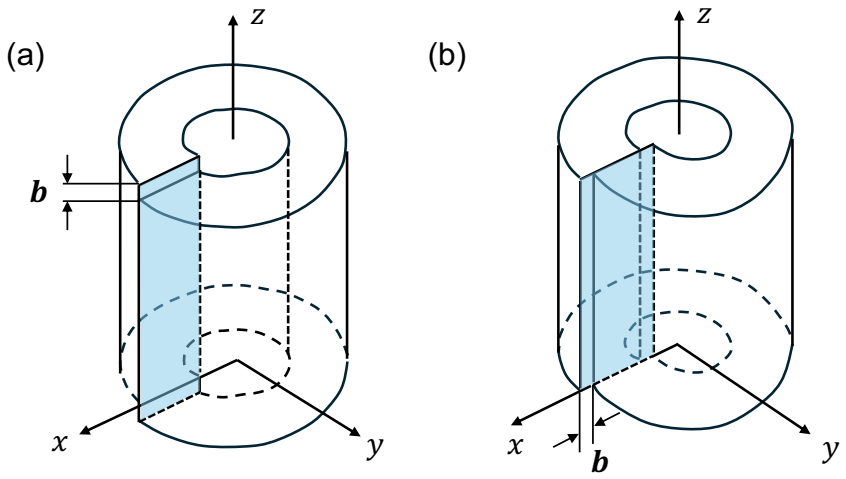

Figure 2.3. Elastic distortion induced by dislocations in cylindrical geometries. (a) Distortion caused by a screw dislocation. (b) Distortion caused by an edge dislocation. The slip plane is shown in blue (adapted from Hull et al. [12]).

$$
\begin{aligned}
\sigma_{xx} &= -D\,y\,\frac{3x^2+y^2}{(x^2+y^2)^2} \\
\sigma_{yy} &= D\,y\,\frac{x^2-y^2}{(x^2+y^2)^2} \\
\sigma_{xy} = \sigma_{yx} &= D\,x\,\frac{x^2-y^2}{(x^2+y^2)^2} \\
\sigma_{zz} &= \nu\big(\sigma_{xx}+\sigma_{yy}\big) \\
\sigma_{xz} = \sigma_{zx} &= \sigma_{yz} = \sigma_{zy} = 0
\end{aligned}
\tag{2.2}
$$

where $D = Gb/2\pi(1-\nu)$. Figure 2.4 shows the stress distribution for infinite straight dislocations. For edge dislocations, the stress field is symmetric and extends both radially and tangentially to the dislocation line. For screw dislocations, the stress field is symmetric only in the radial direction.

The elastic stress fields associated with dislocations are therefore not only fundamental to understanding their individual behavior but also act as the underlying source of the forces driving their interactions. These internal stress fields, together with externally applied stresses, give rise to forces that govern dislocation motion and interactions, which are discussed in the following section.

2.1.3. Forces on Dislocations

The elastic stress fields generated by dislocations, together with externally applied stresses, exert forces that control their motion, interaction, and evolution within a

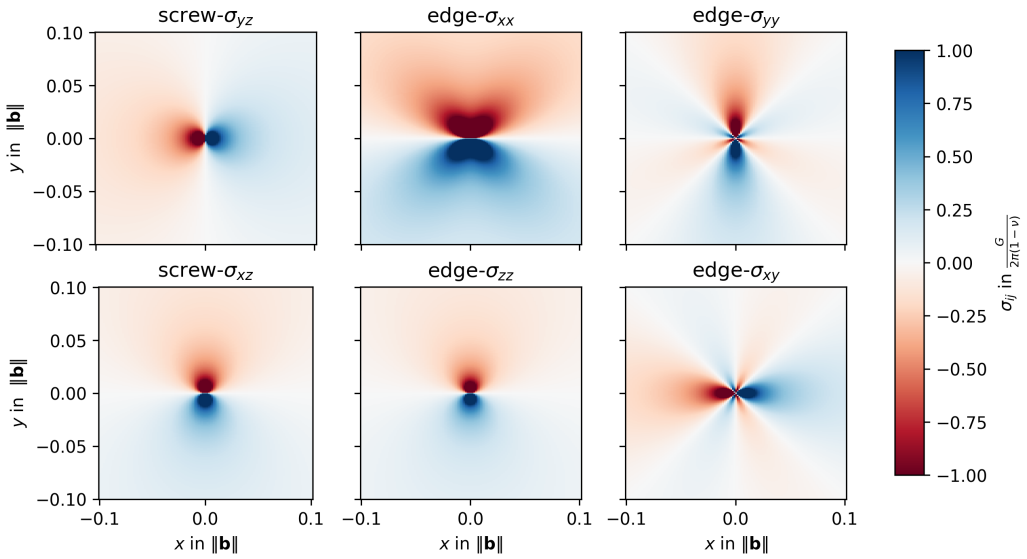

Figure 2.4. Stress field of infinite straight dislocations in an infinite medium.

crystal lattice. Their combined action determines how dislocations move, multiply, and interact, thereby shaping the mechanical response of crystalline materials. Here are some of the forces acting on dislocations:

(1) **Self-energy of a dislocation line**

The presence of dislocations induces local strains and distortions within the lattice, leading to an increase in the overall energy. This part of energy is commonly referred to as strain energy. It consists of two components, namely the core energy (E_c) at close proximity to the dislocation core, and the elastic strain energy (E_{el}) due to the stress field induced by the dislocation. Since the former energy accounts for about 1/10–1/15 of the total strain energy, it is often neglected. In contrast, elastic forces outside the core tend to push or pull the dislocation, attempting to minimize the strain energy. Adopting the continuum medium elasticity model, the expressions for the strain energy, denoted as E_{el}, per unit length of an edge dislocation and a screw dislocation can be obtained as

$$E_{el}(\text{screw}) = \frac{Gb^2}{4\pi}\int_{r_0}^{R_{ext}}\frac{dr}{r} = \frac{Gb^2}{4\pi}\ln\left(\frac{R_{ext}}{r_0}\right) \tag{2.3}$$

$$E_{el}(\text{edge}) = \frac{Gb^2}{4\pi(1-\nu)}\int_{r_0}^{R_{ext}}\frac{dr}{r} = \frac{Gb^2}{4\pi(1-\nu)}\ln\left(\frac{R_{ext}}{r_0}\right) \tag{2.4}$$

where R_{ext} is the outer cutoff radius (i.e., the distance from the dislocation core to the boundary of the region considered for the elastic field) and r_0 is the core radius.

(2) **Line tension**

Dislocations can be regarded as lines of atomic misfit within the crystal lattice, whose total strain energy is proportional to the length of the dislocation line. This results in a so-called line tension, which tends to straighten or shorten the dislocation line in order to minimize its total energy. When a dislocation segment is subjected to a shear stress τ and bent to a radius of curvature R_c, the line tension T exerts a restoring force directed toward the center of curvature, as illustrated in Figure 2.5a. Considering a dislocation segment of length $\mathrm{d}l$, the force per unit length due to the applied shear is τb. The equilibrium condition can thus be written as

$$\tau b \, \mathrm{d}l = 2T \sin\left(\tfrac{\mathrm{d}\theta}{2}\right). \tag{2.5}$$

In the limit $\mathrm{d}\theta \to 0$, this reduces to

$$\tau b = \frac{T}{R_c}. \tag{2.6}$$

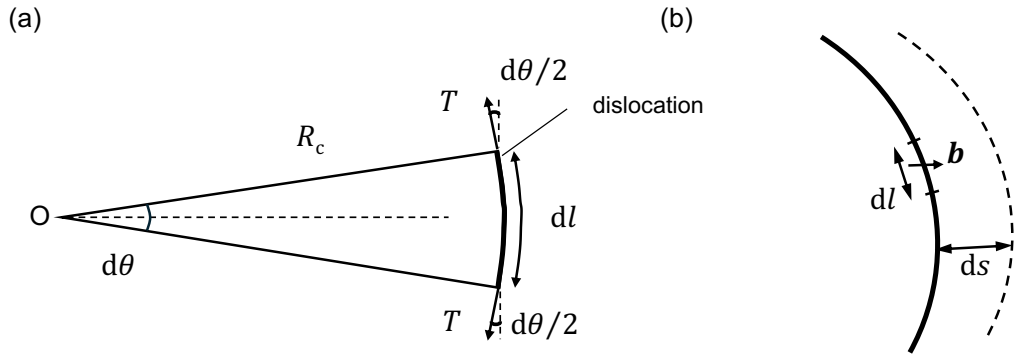

Figure 2.5. Line tension and glide force of a dislocation. (a) Schematic illustration of the line tension of a dislocation. (b) Displacement $\mathrm{d}s$ used to determine the glide force on an element $\mathrm{d}l$ in its slip plane (adapted from Hull et al. [12]).

Given the geometric relation $\mathrm{d}l = R_c \, \mathrm{d}\theta$, it follows that when $\mathrm{d}\theta$ is sufficiently small, $\sin(\mathrm{d}\theta/2) \approx (\mathrm{d}\theta/2)$. Under this small-angle approximation,

$$\tau = \frac{G\,b}{2\,R_c}. \tag{2.7}$$

In other words, a dislocation with both ends pinned will bend into a curved shape with radius of curvature R_c under the action of a shear stress τ.

(3) **Forces on dislocations**

When subjected to an external shear stress τ, dislocations glide along their slip planes in the direction of their Burgers vector $\boldsymbol{b}$. During glide, the motion of a dislocation segment is confined to the slip plane and is typically perpendicular to the dislocation line. This motion can be interpreted as being driven by a force acting on the dislocation segment within the slip plane. According to the principle of virtual work, consider a dislocation segment of length $\mathrm{d}l$ that moves by a distance $\mathrm{d}s$ under the applied shear stress τ. As a result, the crystal slips by the Burgers vector $\boldsymbol{b}$ along the slip plane (see Figure 2.5b). The work done by the applied stress is then

$$\mathrm{d}W = \tau\, b\, \mathrm{d}l\, \mathrm{d}s \tag{2.8}$$

where $b = \|\boldsymbol{b}\|$. This work is equivalent to the work done by the total force F acting on the dislocation segment, which moves through $\mathrm{d}s$,

$$\mathrm{d}W = F\, \mathrm{d}s. \tag{2.9}$$

Hence,

$$F = \tau\, b\, \mathrm{d}l, \tag{2.10}$$

and the force per unit length of dislocation line is

$$f = \frac{F}{\mathrm{d}l} = \tau\, b. \tag{2.11}$$

Here, f denotes the Peach–Koehler force per unit length, which is proportional to the applied shear stress and the Burgers vector of the dislocation. Its direction is perpendicular to the dislocation line and lies in the slip plane.

(4) **Forces between dislocations**

Dislocations can interact with each other, leading to forces between them. These interactions can be attractive or repulsive, depending on the type and relative orientations of the dislocations. The forces arising from dislocation–dislocation interactions can influence the motion, annihilation, or creation of dislocations in a material.

The forces described above ultimately govern the collective behavior of dislocations under load, which in turn determines the onset of yielding in crystalline materials. To understand how these microscopic mechanisms manifest as macroscopic strength, we next examine the relationship between dislocations and yield strength.

2.1.4. Relation between Dislocation and Yield Strength

Since plastic deformation is governed by the motion of dislocations, the yield strength of a material is fundamentally determined by obstacles that impede or modify their

motion [12]. A variety of microstructural features can serve as such obstacles, including grain boundaries, solute atoms, second-phase particles, other dislocations, and even phase transformations. These contributions are often interdependent and can act synergistically, leading to enhanced mechanical properties. The cumulative effect of these mechanisms on the shear yield strength, denoted as τ_y, can be approximated as

$$\tau_y = \tau_i + \tau_{ss} + \tau_{ppt} + \tau_{wh} + \tau_{gb}. \tag{2.12}$$

Here, τ_i denotes the intrinsic lattice resistance to dislocation motion, while τ_{ss}, τ_{ppt}, and τ_{wh} correspond to strengthening by solid solution, precipitates, and dislocation interactions, respectively. In materials with very fine grains, grain boundary strengthening τ_{gb} can make an additional significant contribution by impeding slip across grains. The specific mechanisms that dominate in a given alloy depend on its chemical composition, processing history, and intended application. By tailoring these microstructural features, engineers and materials scientists can design alloys with optimized combinations of strength, hardness, and durability for demanding applications in aerospace, automotive, and structural engineering. The following section focuses on their implications for high-entropy alloys, with particular attention to the CoCrFeMnNi (Cantor) alloy studied in this work.

2.1.5. Strengthening Mechanism in High-Entropy Alloys

While the classical strengthening mechanisms described above are well established in conventional alloys, their manifestations become more complex in multi-component complex alloys (MCCA), such as high-entropy alloys (HEAs). Interest in MCCA has surged due to their remarkable mechanical properties, which often surpass those of conventional single-phase alloys [51, 52, 53, 54, 55]. These alloys, composed of near-equiatomic proportions of four or more elements, exhibit a combination of unique features, including severe lattice distortion, high configurational entropy, and sluggish diffusion, all of which fundamentally alter deformation behavior. Such characteristics can boost conventional strengthening mechanisms and give rise to additional strengthening pathways, thereby contributing to their outstanding mechanical performance.

Among these systems, the CoCrFeMnNi alloy, commonly referred to as the "Cantor" alloy [52], has been studied extensively as a model HEA due to its single-phase face-centered-cubic (fcc) structure and exceptional mechanical response [56, 57]. Notably, it combines high strength with excellent toughness, properties that are retained even at cryogenic temperatures [58, 59]. Figure 2.6 shows representative tensile stress-strain curves of the equiatomic CoCrFeMnNi alloy, demonstrating that this balance of strength and ductility persists at low temperatures, whereas both yield strength and strain-hardening capacity decrease gradually with increasing temperature.

The presence of significant lattice distortion and the complex interactions among multiple constituent elements create a highly intricate atomic and energy landscape for dislocations. This gives rise to an effective form of friction that differs from conventional Peierls-type lattice resistance. To better understand the nature and implications of this

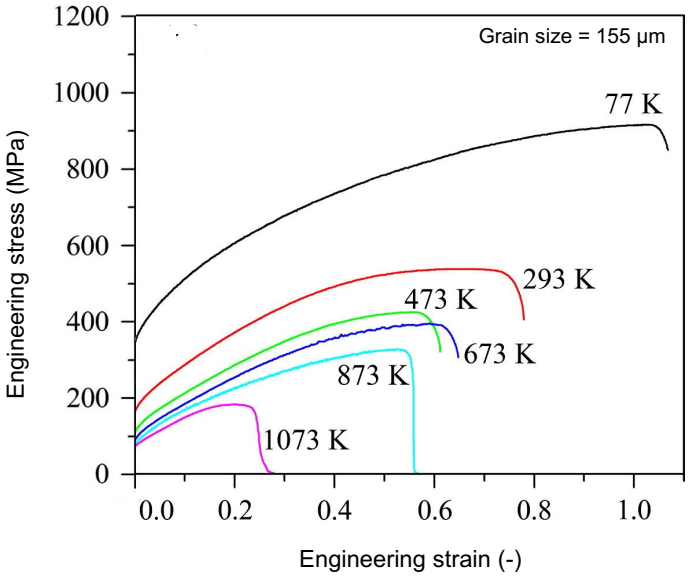

Figure 2.6. Representative engineering stress-strain curve of the equiatomic CoCrFeMnNi high-entropy alloy with a grain size of 155 µm, tested at different temperatures (adapted from Otto [59]). The alloy exhibits simultaneously high strength and ductility at cryogenic temperature (77 K).

frictional resistance, numerous studies have investigated the underlying strengthening mechanisms in HEAs.

Wu et al. [60] conducted an investigation on the thermal activation processes in equiatomic fcc single phase solid-solution alloys. They examined stress-strain measurements at different temperatures and strain rates. Based on their findings, it was concluded that the deformation behavior in equiatomic alloys is primarily governed by the Labusch-type solution strengthening mechanism, rather than lattice friction. The Labusch model [61] provides a classical description of solid-solution strengthening, in which dislocations interact collectively with statistically distributed solute atoms in concentrated alloys. In the context of high-entropy alloys (HEAs), solute strengthening is believed to be associated with the occurrence of either short-range clustering or short-range ordering of solute atoms [62, 63, 64, 65]. However, Lee et al. [66] did not detect any chemical heterogeneity or ordered structures in the CoCrFeMnNi alloy, even when employing experimental techniques of sizes down to the resolution limits of 1 nm, such as electron diffraction, scanning transmission electron microscopy with energy-dispersive X-ray spectroscopy (STEM-EDS) and atom probe tomography (APT) at cryogenic temperatures.

Certain researchers suggest that the phenomenon of strengthening is attributed to phase transformation, as theoretical models based on first-principles predicted the existence of a stable hexagonal close-packed (hcp) phase at lower temperatures [67, 68]. Among these, Yu et al. [67] simulated that the proportion of the stronger hcp phase gradually increases as plastic deformation occurs within the fcc phase in a

Cantor-like $Cr_{20}Mn_6Fe_{34}Co_{34}Ni_6$ alloy. However, the hcp phase has not been found in the equiatomic Cantor alloy for the time being [66].

More people have recently recognized that local chemical fluctuation is a common feature in HEAs and is believed to have an impact on the motion of dislocations [69, 70, 71, 72]. Curtin et al. [69] have proposed a theoretical framework aimed at explaining the plastic yield strength observed in fcc high entropy alloy (HEA)s. This theory has been justified using molecular simulations conducted on Fe-Ni-Cr alloy models. The authors regard each individual elemental components as a solute embedded inside the effective matrix of the surrounding alloy. They have inferred that the primary mechanism by which strengthening occurs is through dislocation interactions with the stochastic local concentration fluctuations present around the average composition. In another study, Ma et al. [70] developed an atomic interaction potential for the CrCoNi medium-entropy alloy. Their findings revealed that the local chemical ordering underwent changes throughout the processing stages, resulting in an enhanced roughness of the local energy landscape. Additionally, this led to an increase in the activation barriers that govern dislocation activities.

The aforementioned studies primarily focus on theoretical aspects and propose that the hindered movement of dislocations on their gliding planes in HEA is primarily attributed to the pinning points. These pinning points arise from chemical and physical distortions within the lattice and possess the potential to exhibit a high density, thereby generating a significant friction force that affects all dislocation movements.

However, there are currently very few quantitative reports on pinning effects, because the interplay between energy and the very local chemical ordering is difficult to characterize and measure. Besides, small-scale simulations are spatially and temporally constrained, which does not allow them to grab a realistic and statistical view of such obstacles and their potential strength. Since moving dislocations carry a shear deformation on the order of the interatomic distance, the energy landscape experienced by one dislocation may differ substantially from that encountered by subsequent dislocations. This is especially true in alloys where local order may be built up or inversely deconstructed by successive shears of dislocations gliding on the same plane. As so, a moving dislocation, if one can follow its individual or collective motion, is the ultimate probe to explore such distorted landscapes as well as the vector of the on-going plastic deformation.

While the preceding sections have examined the fundamental mechanisms of dislocation-mediated plasticity, direct experimental quantification of these microscopic processes remains highly challenging. Establishing quantitative links between microstructural evolution and macroscopic mechanical behavior requires not only advanced characterization techniques but also careful observation of dislocation dynamics over relevant spatial and temporal scales. Here, the first part of this work aims to characterize and quantify the nature of the pinning points encountered by moving dislocations over extended spatial and temporal scales, based on experimental observations obtained from *in situ* TEM straining.

2.2. Development of Hardness Measurement and Nanoindentation

Hardness testing is one of the most widely used techniques for assessing the mechanical response of materials, providing a convenient measure of their resistance to plastic deformation. Traditional hardness testing methods, such as Brinell, Vickers, or Rockwell testing, operate at the macro- or microscale and are well suited for evaluating bulk mechanical properties.

With advances in instrumentation and experimental control, these techniques have evolved toward the nanoscale, giving rise to nanoindentation—a highly sensitive extension of hardness testing that enables the extraction of mechanical properties from extremely small volumes of material. In addition to hardness, nanoindentation can provide elastic modulus, time-dependent deformation behavior, and microstructure-sensitive mechanical signatures, making it a powerful tool for modern materials characterization.

2.2.1. Classical Hardness Theory and Contact Mechanics

The theoretical foundation of nanoindentation builds upon classical contact mechanics and hardness theory, which together provide the fundamental link between applied load, deformation response, and material strength [73, 74]. Central to this theoretical framework is the elastic contact theory formulated by Heinrich Hertz in 1881 [75], which describes the stress distribution, displacement field, and contact area when two isotropic, linearly elastic bodies come into frictionless contact. Although Hertz's model is limited by its assumption of purely elastic behavior, its analytical expressions for contact stiffness and contact area remain essential to modern indentation analysis.

Building on this mechanical framework, macroscopic hardness testing methods were developed in the early 20th century to quantify a material's resistance to plastic deformation. The Brinell hardness test [76] uses a steel ball indenter to press into the material surface (see Figure 2.7a) and derives hardness from the measured diameter of the indentation area, making it suited for relatively soft metals. The Vickers test [77], which employs a pyramidal indenter (see Figure 2.7b), later became the precursor to microhardness testing owing to its versatility and the self-similar nature of its indentation geometry. While these approaches enabled robust characterization of bulk materials, their spatial resolution and precision were insufficient for thin films, heterogeneous microstructures, and nanoscale materials [78].

It is worth noting that the understanding of strength mechanisms during this period remained rudimentary. Griffith's energetic criterion of fracture mechanics [80] in 1920, and the concurrent introduction of dislocations by Orowan, Polanyi, and Taylor in the 1930s laid the groundwork for modern plasticity theory, connecting macroscopic mechanical response to underlying defect processes. This conceptual link was made quantitative in Tabor's classic monograph *The Hardness of Metals* in 1951 [81], which established a direct relationship between indentation hardness

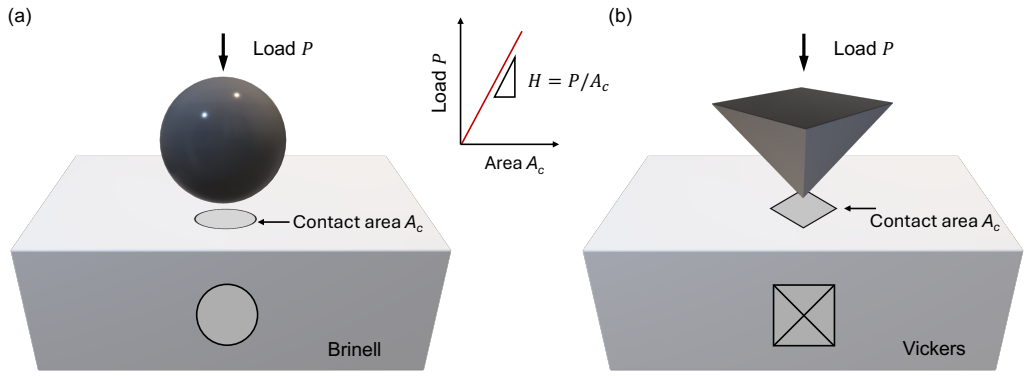

Figure 2.7. Schematic comparison of indentation geometry. (a) Spherical indenter (Brinell). (b) Diamond pyramid (Vickers) (adapted from Hutchings [79]). The indent indicates plastic deformation, and the material's strength is defined as the load P divided by the projected contact area A_c perpendicular to the load.

and yield strength ($H \approx 3\sigma_y$). Analytical advances followed, including Sneddon's solution [82] for axisymmetric indenters on elastic half-spaces, which provided the mathematical foundation for interpreting unloading curves in modern nanoindentation experiments.

2.2.2. The Oliver–Pharr Model and Load–Displacement Analysis

The establishment of nanoindentation as a systematic and quantifiable technique was significantly advanced by the work of Oliver and Pharr in 1992 [83]. Their model offered a robust methodology for extracting elastic modulus (E) and hardness (H) directly from the load–displacement ($P - h$) curve, eliminating the dependence on optical measurements of indentation dimensions in traditional methods and enabling the characterization of mechanical properties at the nanoscale. To date, the Oliver–Pharr method remains the cornerstone of instrumented nanoindentation data analysis.

Unlike earlier approaches that focused solely on the loading phase, this method is centered on analyzing the unloading segment of the curve. The model rests on three key assumptions: (1) the material exhibits primarily elastic recovery during unloading, (2) the indenter is a rigid body with known geometry (e.g., Berkovich, Vickers, or conical), and (3) the loading-unloading cycle is stable and reproducible, free from significant time-dependent effects such as creep.

The core idea is to extract the contact stiffness S from the slope at the maximum load point ($h = h_{\max}$, $P = P_{\max}$) of the unloading curve,

$$S = \left.\frac{\mathrm{d}P}{\mathrm{d}h}\right|_{h=h_{\max}}.$$

This enables the estimation of the contact depth h_c as

$$h_c = h_{max} - \epsilon \cdot \frac{P_{max}}{S},$$

where ϵ is a geometric constant (typically 0.75 for Berkovich indenters). From this, the projected contact area A_c as a function of h_c is calculated (e.g., ideal Berkovich: $A_c(h_c) = 24.5{h_c}^2$), leading to the determination of two fundamental mechanical quantities hardness (H) and the reduced modulus(E_r),

$$H = \frac{P_{max}}{A_c},$$

$$E_r = \frac{S}{2} \cdot \sqrt{\frac{\pi}{A_c}}.$$

The corresponding hardness H describes the material's resistance to plastic flow, whereas the slope of the unloading curve $\mathrm{d}P/\mathrm{d}h$ reflects the contact stiffness S, from which the elastic modulus is derived. Together, they provide complementary measures of a material's mechanical response under indentation. Finally, using the relation between sample and indenter properties,

$$\frac{1}{E_r} = \frac{1-\nu^2}{E} + \frac{1-{\nu_i}^2}{E_i},$$

one obtains the elastic modulus E of the test material, where ν and E are the Poisson's ratio and elastic modulus of the sample, and ν_i, E_i correspond to the indenter.

The analysis process is highly dependent on the load–displacement ($P-h$) curve. A typical $P-h$ curve consists of three regions: the loading phase representing elastoplastic deformation, a (potential) hold segment used to assess creep, and the unloading phase where elastic recovery dominates. These regions, along with parameters such as P_{max}, h_{max}, and S, are illustrated in Figure 2.8.

One of the key advantages of the Oliver–Pharr method is that it does not require direct measurement of the contact area, which makes it especially suitable for micro-scale testing—a critical benefit in nanoscale experiments where optical imaging is difficult or even impossible. Hence, the Oliver–Pharr framework has become the modern standard for interpreting nanoindentation results and is widely applied to the characterization of thin films, multilayers, and materials with depth-dependent mechanical behavior.

2.2.3. Applications and Challenges of Nanoindentation

The nanoindentation analysis method proposed by Oliver and Pharr [78, 84] has since been widely adopted and continuously developed over the past decades. With increasing complexity in material and advancements in experimental precision, researchers have sought to extend indentation analysis into more complex physical scenarios [78]. For

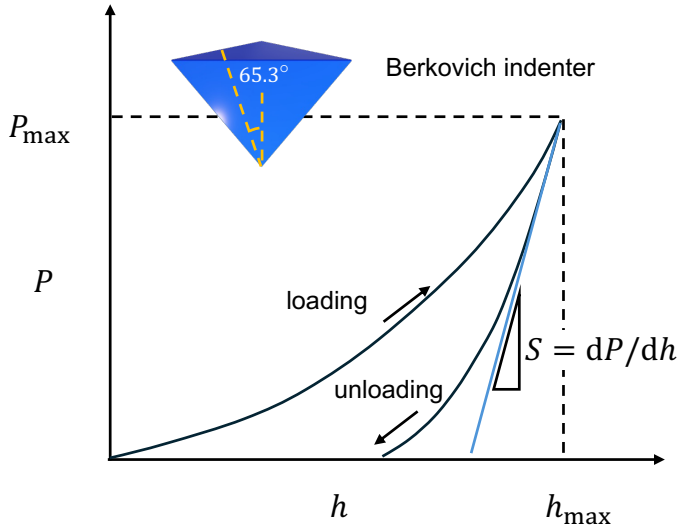

Figure 2.8. Representative load–displacement curve, illustrating the loading, unloading and slope of unloading curve (dP/dh).

example, the Nix-Gao model explained the indentation size effect (ISE), where hardness increases significantly at shallow depths due to geometrically necessary dislocations [85]. Despite the development of nanoindentation theory, these approaches share a common premise: they are primarily based on the mechanical response of a single or limited number of indentation points, making them suitable for localized interpretation of material behavior. However, in many complex materials, results derived from isolated indents often fail to capture the microstructural details of complex, heterogeneous, or multiphase materials.

To this end, statistical nanoindentation become increasingly important. Unlike conventional indentation techniques, statistical nanoindentation employs large-scale grid indentations across the sample surface to acquire extensive datasets of local hardness and elastic modulus [86, 87, 88], as indicated in Figure 2.9. These datasets reveal the spatial distribution of mechanical properties at the microscale. For instance, studies have reported approximately 500,000 indentations on eutectic Al–Cu alloys and duplex stainless steels [89], and 100,000 and 212,500 indentations on the Taza meteorite [90] and thermal-barrier coatings [91], respectively. Such large datasets can be used not only to calculate average mechanical properties but also to map spatial distributions [92, 93], identify microstructural boundaries, and analyze the correlation between microstructure and mechanical properties [93, 94, 95].

The massive volume of such data has driven the adoption of advanced data science techniques [97]. Traditional statistical or physical models often struggle with this scale and complexity. Recent studies have thus integrated machine learning approaches for the analysis and interpretation of nanoindentation data. Studies include the use of Gaussian mixture model (GMM) and the expectation-maximization (EM) algorithm [98] for decomposing multimodal data distributions [99, 100]; convolutional neural network (CNN) for identifying "pop-in" events in the load–displacement curves

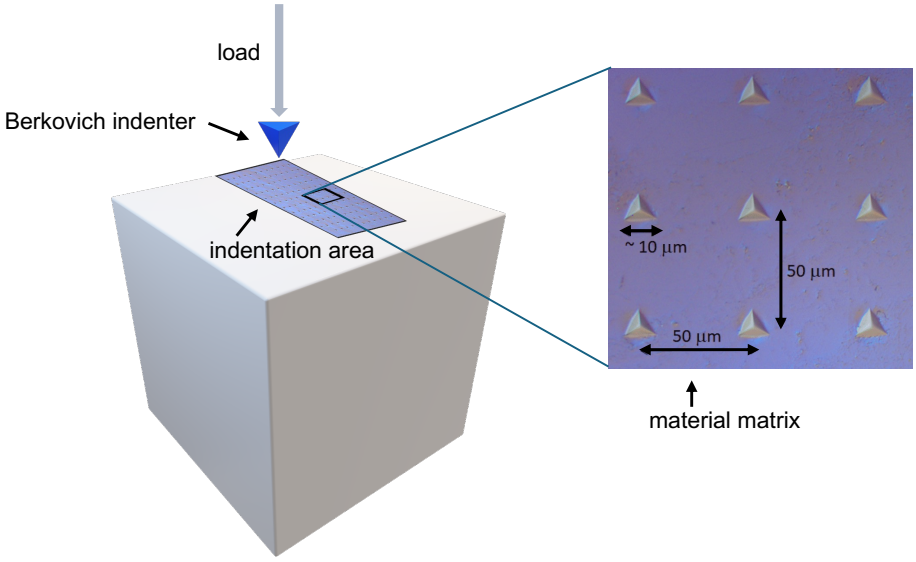

Figure 2.9. Schematic diagram of the nanoindentation grid. Experimental image adapted from Karanja [96]: The space between indents was set to 50 µm. The indentation size was about 5 µm-10 µm.

[101]; k-means clustering to extract phase level features, as applied to duplex and high-strength low-alloy steels [102]; and Bayesian inference methods for estimating material properties and uncertainties, providing a more robust and comprehensive analysis of nanoindentation data [97, 103]. These studies show that incorporating machine learning into nanoindentation analysis not only enhances the resolution of microstructural features, but also extends the scale of mechanical characterization, representing a key trend in future materials characterization.

This study is situated within this context and aims to explore how machine learning techniques can be used to model and analyze large-scale load–displacement data from nanoindentation experiments, in order to reveal the underlying links between microstructure and mechanical properties. Additionally, a cross-validation approach was introduced to assess whether the available data quantity was sufficient and to estimate the amount of data required for reliable results, one of the frequently encountered yet difficult-to-address challenges in applying ML to materials science problems.

2.3. Materials Data Science and Informatics

The emergence of data science and artificial intelligence has significantly reshaped the landscape of materials research [104]. While traditional experimental techniques have made significant progress in characterizing microstructures, they often fall short when dealing with complex phenomena, such as internal stress distributions, or high-dimensional structure-property relationships. As highlighted in the first chapter, these limitations call for a more scalable, integrative, and intelligent approach to materials

characterization. Materials data science, situated at the intersection of materials science, computer science, and statistics, provides such a framework [3].

This section introduces the core objective of materials data science, the infrastructure enabling large-scale data management and analysis, and the recent advances in data-driven microstructure characterization. These elements collectively underpin the methodologies developed in this thesis.

2.3.1. The Main Objective of Materials Data Science

The main objective of materials data science is to systematically establish and exploit the intrinsic links between processing conditions, microstructural features, and macroscopic material properties, a concept often referred to as the processing-structure-property (PSP) relationship [105]. This relationship has always been fundamental to materials science. To this end, materials data science offers a paradigm shift, reframing PSP modeling as a data-driven inference task. This enables researchers to extract statistically meaningful correlations, discover latent patterns, and build predictive models capable of generalizing across material systems. In this context, data is not merely a byproduct of experimentation or simulation, it becomes a central asset, actively shaping the trajectory of material discovery and design. As shown in Figure 2.10, this asset progresses from information to materials knowledge and ultimately to wisdom, enabling customized materials design for specific applications.

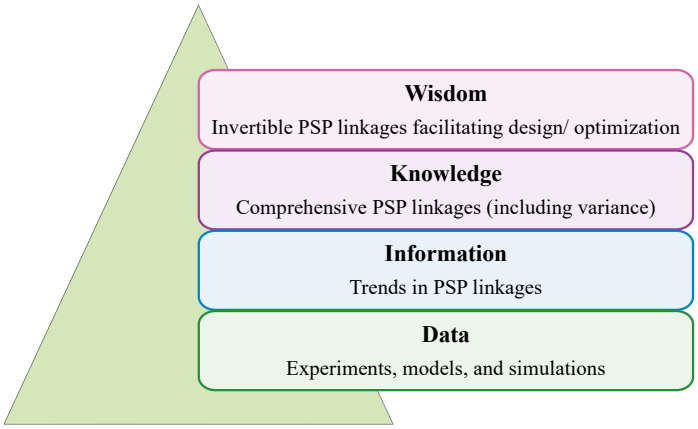

Figure 2.10. Hierarchical scheme for the main objectives of materials data science, adapted from Kalidindi [104].

2.3.2. Infrastructure for Data-Driven Materials Science

The success of data-driven approaches in materials research depends not only on algorithms and data availability, but also on a well-established infrastructure that supports the entire data life cycle, from acquisition to interpretation and collaborative use. As emphasized in initiatives such as the MGI, building a robust digital ecosystem is

essential. This ecosystem is typically structured around three foundational pillars [104]: Data management, data analytics, and electronic collaboration science.

(a) Data Management: Data management refers to the processes of organizing, storing, annotating, and standardizing materials data to ensure that it is findable, accessible, interoperable, and reusable (FAIR principles [106]). Given the diversity of data sources, such as microscopy images, spectroscopic signals, simulation outputs, and mechanical test results, maintaining consistency in formats, metadata annotation, and version control is critical. Electronic Laboratory Notebooks (ELNs) [107], cloud-based data repositories [108], and structured databases such as the Materials Project and NOMAD provide necessary infrastructure to manage data at scale, enabling traceability and reproducibility across research groups.

(b) Data Analytics: Once data is curated, the next step is to extract knowledge from it. Data analytics involves the application of statistical methods, machine learning algorithms, and visualization tools to uncover hidden relationships, identify patterns, and construct predictive models. In the context of materials science, data analytics has enabled property prediction, high-throughput screening, and even inverse materials design [109].

(c) Electronic Collaboration Science (e-CS): Modern materials science is inherently interdisciplinary and collaborative. The e-CS refers to the platforms and tools that enable researchers to share data, models, and workflows seamlessly across institutional and national boundaries. Platforms such as Citrination [110], AFLOW [25], and the open knowledgebase of interatomic models (OpenKIM) [111] project exemplify this trend, providing open-access APIs, shared repositories, and community-driven model libraries. These platforms not only reduce duplication of effort but also accelerate innovation by promoting transparency and re-usability of computational and experimental results.

These three pillars form the backbone of the data-driven infrastructure that supports modern materials informatics. They enable researchers to efficiently manage complex datasets, extract actionable insights, and collaborate across disciplines and institutions. In this thesis, the proposed methodologies for microstructure characterization are designed to integrate into this infrastructure, ensuring scalability, transparency, and interoperability with future materials research platforms.

2.3.3. Applications in Materials Characterization

In data-driven materials science, characterization has progressed from qualitative description toward quantitative, scalable, and increasingly automated analysis supported by advanced computational techniques [112, 113]. This section reviews recent advances in applying data science to experimental materials characterization, with a focus on microstructural features relevant to this thesis:

- **Image-based microstructure quantification:** Extracting microstructural features from microscopy images remains a central challenge, particularly under imaging conditions of noise, overlap, or complex morphologies. Recent data-driven methods have enabled automated and accurate quantification.

  For example, Bilyk et al. [114] developed a Mask R-CNN-based framework for TEM images of irradiated HEAs, achieving automated segmentation and quantification of thousands of dislocation loops with sub-percent error, even under high-overlap conditions. Meanwhile, Govind et al. [115] introduced a deep learning approach that leverages synthetic training data and pretraining, effectively addressing the problem of limited annotated TEM images for defect analysis. In another study, DeCost et al. [116] applied texture descriptors and dimensionality reduction to ultrahigh carbon steels, demonstrating how microstructures can be mapped into interpretable low-dimensional manifolds. These studies establish image-based ML as a accurate and scalable tool for microstructural quantification.

- **Integration with experimental techniques:** Variability in measurement protocols and analysis methods complicates the integration of data across experimental domains. Recent work shows that machine learning can improve the robustness of nanoindentation and other mechanical testing workflows, thereby supporting more reproducible characterization.

  Puchi-Cabrera et al. [97] reviewed the use of machine learning in nanoindentation, covering methods for curve classification, property prediction, and feature extraction, and highlighting how reproducibility and standardization remain essential for translating indentation data into reliable materials insights. From another perspective, Spurgeon et al. [117] outlined the vision of data-driven transmission electron microscopy, emphasizing the integration of experimental platforms with advanced data analysis to enable real-time, quantitative characterization.

- **Defect detection and dislocation analysis:** One of the core interests of this thesis lies in the detection and analysis of dislocations, which are among the most critical defects governing plasticity. Although their characterization across scales remains difficult, recent data-driven approaches are making defect analysis more quantitative and interpretable.

  Hochrainer et al. [118] present a novel plasticity theory based on systematic physical averages of the kinematics and dynamics of dislocation systems that can predict microstructure evolution and size effects in accordance with experiments and discrete dislocation simulations. Building on this foundation, Steinberger et al. [119] used machine-learning classifiers to categorize dislocation microstructures generated by discrete dislocation dynamics simulations, demonstrating how data-driven models can capture and distinguish characteristic patterns in dislocation structures. Extending these approaches to experimental data, Song et al. [120] developed a deep learning-based, high-throughput pipeline tailored to *in situ* TEM, enabling quantitative tracking of dislocation motion and revealing avalanche-like dynamics at

unprecedented temporal resolution. This progression from theory to simulation to experiment underscores the interpretive power of data-driven dislocation analysis.

- **Toward intelligent, automated characterization:** Balancing data availability and model performance is a persistent challenge, especially when data acquisition is costly and time-consuming. A promising direction is the integration of machine learning with autonomous experimental platforms.

  A compelling example is provided by Wagner et al. [121], who introduced the autonomous materials and device application platform (AMANDA) - an integrated system of robotics, software orchestration, and closed-loop experimental control for facilities like organic solar cell screening. The embodiment LineOne is able to autonomously prepare and characterize hundreds of solution-processed thin-film devices per day, complete with full data tracing and adaptive experiment planning. This work exemplifies how intelligent, automated workflows can simultaneously enhance throughput, reproducibility, and experimental efficiency-effectively addressing the challenge of limited data and high experimental costs.

In summary, data-driven characterization is reshaping how microstructural information is extracted, analyzed, and interpreted. The four application areas reviewed here not only respond to the challenges identified in Section 1.3 but also define the methodological foundation on which the approaches developed in this thesis are built.

Part II.

METHODS

3. 3D Reconstruction of Dislocations in TEM

To gain a deeper understanding of the strengthening mechanisms governed by dislocation motion, it is essential to analyze dislocation activity on specific slip systems. Such analysis requires not only the identification of individual dislocations but also their accurate spatial representation. The three-dimensional geometric reconstruction of dislocations has therefore become an important part of TEM studies, where the reliability of such reconstruction strongly depends on the accurate determination of experimental projection geometry. This chapter addresses the methodological considerations for three-dimensional representation of dislocations, with particular emphasis on achieving projection accuracy in TEM experiments.

3.1. Orientation and Coordinate Systems

3.1.1. Orientation of Individual Crystallites

A material composed of multiple crystallites is called a polycrystalline material, and the individual crystallites are oriented differently within the sample. In the case of sheet and foil, the orientation refers to the relationship between the crystal coordinate system (CCS) along the [100]-[010]-[001] directions and the sample coordinate system (SCS) along the [rolling direction (RD)]-[transverse direction (TD)]-[normal direction (ND)] directions, as shown in Figure 3.1.

To depict the rotation of two coordinate systems with respect to each other, or to describe the orientation of a crystal, there exist various methods, such as Euler angles, rotation quaternions, and others. These representations are essentially equivalent and can be interconverted. Nonetheless, specific purposes may favor certain representations, and different "communities" may adopt distinct conventions—sometimes without explicitly stating them. This has led to frequent inconsistency and confusion in the representation and conversion of 3D rotations [123].

3.1.2. Reference Coordinate Systems

To unambiguously describe the slip orientations of dislocations within TEM samples, three reference coordinate systems are introduced in this study. All coordinate systems follow the standard right-handed Cartesian convention, in which the right thumb points along the positive z-axis and the curl of the fingers indicates the positive rotation

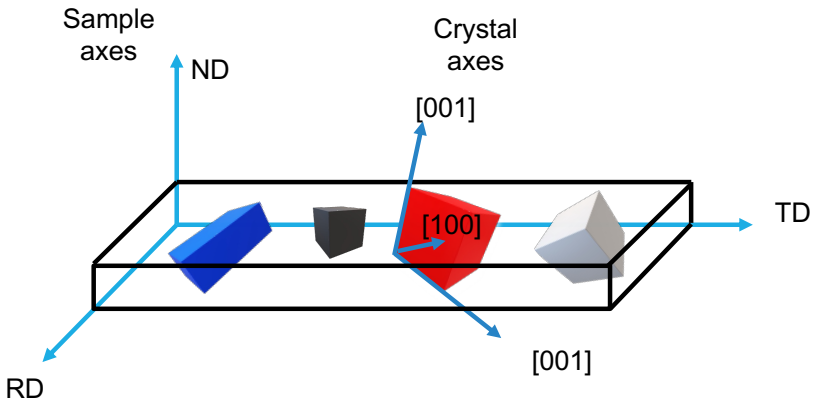

Figure 3.1. Orientation of crystallites in the sample: rolling direction (RD), transverse direction (TD), and normal direction (ND) [122].

from the x-axis toward the y-axis. A rotation about the z-axis is defined as positive if, when looking along the positive z-axis towards the origin, the rotation appears counterclockwise. The three orthogonal axes thus form an orthonormal right-handed basis in 3D space. Based on this convention, three reference coordinate systems are specified:

1. The world coordinate system (WCS) is defined by the experimental setup, with basis vectors $\{\boldsymbol{x}_\mathrm{w}, \boldsymbol{y}_\mathrm{w}, \boldsymbol{z}_\mathrm{w}\}$. It provides a fixed laboratory reference frame.

2. The sample coordinate system (SCS) is defined according to the specimen geometry, with basis vectors $\{\boldsymbol{x}_\mathrm{s}, \boldsymbol{y}_\mathrm{s}, \boldsymbol{z}_\mathrm{s}\}$ typically aligned with the sample edges.

3. The crystal coordinate system (CCS) is established along the crystallographic axes, with basis vectors $\{\boldsymbol{x}_\mathrm{c}, \boldsymbol{y}_\mathrm{c}, \boldsymbol{z}_\mathrm{c}\}$, serving to describe slip directions and planes in the lattice.

In *in situ* transmission electron microscopy (TEM) experiments, the WCS is defined along the laboratory axes and follows the right-handed convention, as illustrated in Figure 3.2. The orientation of the specimen with respect to the incident beam is adjusted by tilting the TEM holder, which allows specific zone axes to be aligned with the beam. Selected area electron diffraction patterns are recorded to confirm that the specimen is oriented close to a desired crystallographic direction. Once an appropriate zone axis is established, real-space imaging is performed under the same orientation to capture dislocation projections. In this study, the specimen is single-tilted by an angle θ around the y_w-axis.

The relationship between the WCS and the SCS is then determined by the specimen tilt angle θ around the y_w-axis. The sign of θ follows the right-handed convention (again): A rotation about the y_w-axis is defined as positive if, when looking in the direction of the negative y_w-axis (i.e. towards the origin), the rotation appears counterclockwise. The orientation of the CCS with respect to the SCS is described

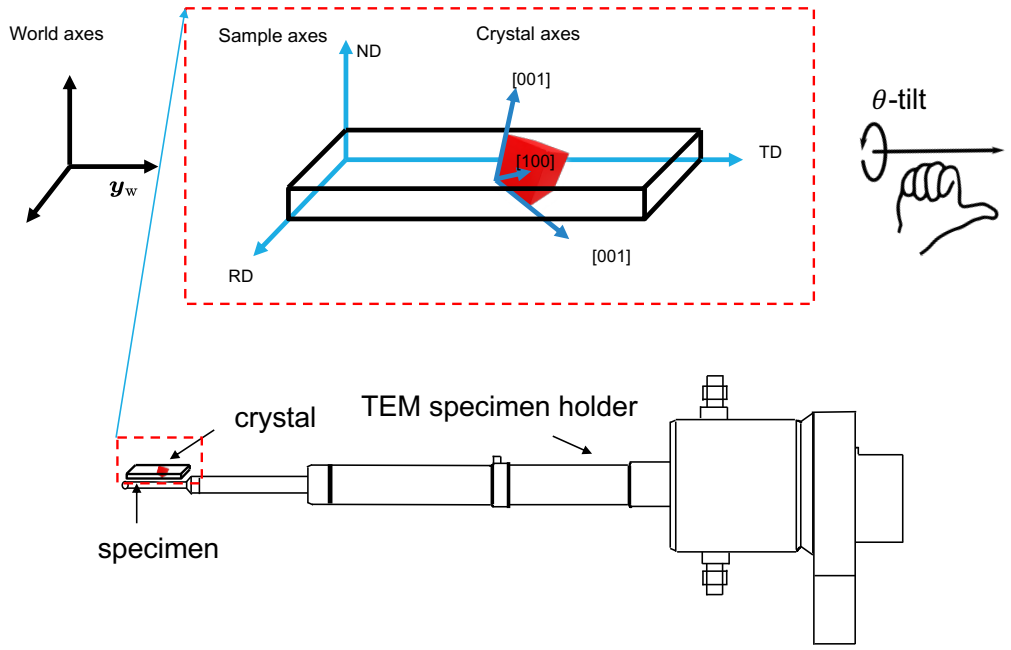

Figure 3.2. Base axes of three coordinate systems, drawn according to the right-hand rule. For clarity, the axis triads are intentionally offset; in practice, the origins of WCS and SCS coincide.

by the Bunge Euler angles (ϕ_1, Φ, ϕ_2), which are introduced here only in notation. Their detailed definition and the corresponding rotation matrix will be presented in the following section, with implementation provided in the associated code repository.

3.2. Chained Rotations

In practical applications, a single rotation is rarely sufficient; instead, multiple rotations are combined in sequence, leading to what is referred to as chained rotations. The interpretation and implementation of chained rotations depend on several conventions. First, one must distinguish between active and passive rotations, which determine whether the object or the coordinate system is considered to be rotating. Second, the order of successive rotations must be specified as either intrinsic (about moving axes) or extrinsic (about fixed axes). Finally, these conventions translate into different rules for matrix multiplication, depending on whether transformations are applied in a pre- or post-multiplication scheme.

3.2.1. Active vs. Passive Rotations

The first distinction concerns whether a rotation is interpreted as acting on the object itself (active) or on the reference frame (passive). Although mathematically equivalent, the two perspectives lead to different physical interpretations and must be clearly

specified in crystallographic applications.

Analytic geometry in 3D Euclidean space distinguishes between active and passive spatial transformations. In passive transformations, the point or vector remains unchanged while the coordinate frame, or basis, is transformed, as shown in Figure 3.3. On the other hand, in active transformations, the vectors are transformed while the coordinate system remains fixed.

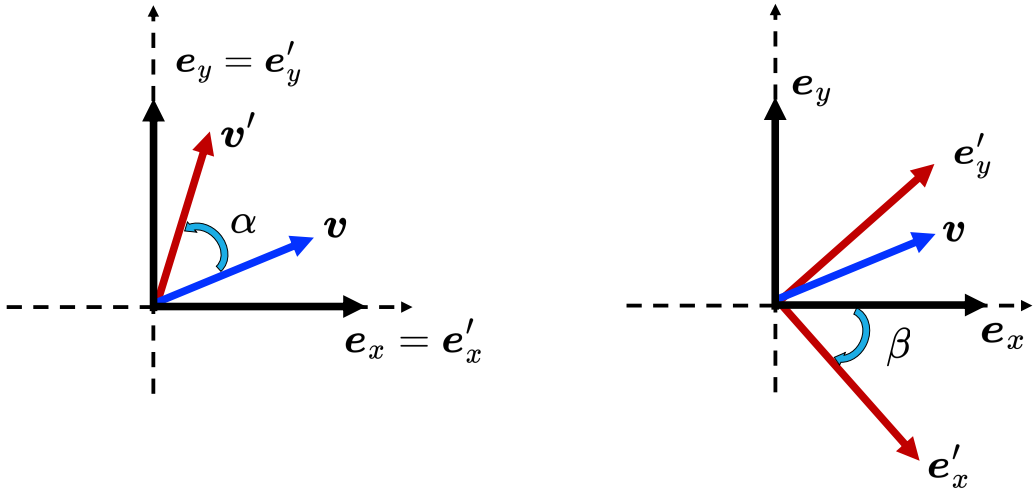

Figure 3.3. Active and passive transformations. (a) Active transformation, in which the object is rotated with respect to a fixed coordinate system. (b) Passive transformation, in which the coordinate system is rotated while the object remains fixed.

The matrix expression for active rotation along the z-axis is given by

$$\mathbf{R}_z^{\mathrm{a}} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow \text{in2D} : \boldsymbol{v}' = \mathbf{R}_z^{\mathrm{a}} \cdot \boldsymbol{v} = \begin{bmatrix} v_x \cos\alpha - v_y \sin\alpha \\ v_x \sin\alpha + v_y \cos\alpha \end{bmatrix} \tag{3.1}$$

where $\boldsymbol{v} = \boldsymbol{v}_x \mathbf{e}_x + \boldsymbol{v}_y \mathbf{e}_y$ and $\boldsymbol{v}'$ is the rotated vector. To put it in a more compact way, the tensor representation is: $\boldsymbol{v}' = \mathbf{R}_z^{\mathrm{a}} \cdot \boldsymbol{v} = \mathbf{R}_z^{\mathrm{a}} \boldsymbol{v}$. The rotated vector $\boldsymbol{v}'$ and the vector $\boldsymbol{v}$ are in the same coordinate space.

The matrix expression for passive rotation along the z-axis is given by

$$\mathbf{R}_z^{\mathrm{p}} = \begin{bmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow \text{in2D} : \boldsymbol{v}' = \mathbf{R}_z^{\mathrm{p}} \cdot \boldsymbol{v} = \begin{bmatrix} v_x \cos\beta + v_y \sin\beta \\ -v_x \sin\beta + v_y \cos\beta \end{bmatrix} \tag{3.2}$$

It should be noted that, in the passive rotation, $\boldsymbol{v}'$ is the new representation of $\boldsymbol{v}$ in the rotated coordinate system. Additionally, it can be observed that two different rotation matrices yield the same result if $\beta = -\alpha$, which means that $\boldsymbol{v}'$ in the world coordinate system appears the same as the original $\boldsymbol{v}$ when observed in the rotated system.

Given the rotation matrix, a real square matrix whose columns and rows are orthonormal vectors (an orthogonal matrix), this can be expressed mathematically as $\mathbf{R}^\top\mathbf{R} = \mathbf{R}\mathbf{R}^\top = \mathbf{I}$, where $\mathbf{R}^\top$ is the transpose of $\mathbf{R}$ and $\mathbf{I}$ is the identity matrix. This lead to the equivalent characterization: a matrix $\mathbf{R}$ is orthogonal if its transpose is equal to its inverse: $\mathbf{R}^\top = \mathbf{R}^{-1}$, where $\mathbf{R}^{-1}$ is the inverse of $\mathbf{R}$. Based on the definition of active and passive rotation matrices given above, we can conclude that $\mathbf{R}^{\mathrm{a}}(\theta) = \mathbf{R}^{\mathrm{p}}(-\theta) = \left[\mathbf{R}^{\mathrm{p}}(\theta)\right]^{-1} = \left[\mathbf{R}^{\mathrm{p}}(\theta)\right]^\top$. With three elemental rotations (rotations about the axes of a coordinate space), it is always possible to achieve any desired target frame. The three elementary transformation matrices for both active and passive conditions are listed below.

Active:

$$\mathbf{R}_x^{\mathrm{a}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}, \mathbf{R}_y^{\mathrm{a}} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \mathbf{R}_z^{\mathrm{a}} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Passive:

$$\mathbf{R}_x^{\mathrm{p}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}, \mathbf{R}_y^{\mathrm{p}} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}, \mathbf{R}_z^{\mathrm{p}} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In practice, active transformation matrices are more commonly employed, as we are generally accustomed to interpreting rotations from the perspective of the object being transformed rather than from that of a rotating coordinate system. When dealing with crystallite orientations, we must be conscious of the perspective we are employing, particularly when the Bunge Euler angles use the object view. In one line, active/passive rotations concern what is regarded as rotating: the object or the reference frame.

3.2.2. Intrinsic vs. Extrinsic Sequences

In chained rotations, an additional ambiguity arises from how successive rotations are defined and applied. This concerns two distinct but related issues: whether rotations are interpreted as intrinsic or extrinsic sequences, and whether they are implemented through pre- or post-multiplication of matrices. The following two subsections first clarify the conceptual distinction between intrinsic and extrinsic sequences, using the Bunge Euler convention as an illustrative example, and then explain how these conventions translate into explicit matrix operations.

A chained rotation can be described either in an intrinsic or an extrinsic manner. In an intrinsic sequence, each successive rotation is performed about an axis of the *rotating* coordinate system, i.e., the axes themselves move after each operation. In contrast, an extrinsic sequence is defined with respect to the *fixed* initial axes, such that all rotations are performed relative to the same reference frame. Although mathematically equivalent, the two interpretations reflect different physical viewpoints.

This distinction is particularly relevant when using Euler angles to describe crystallite orientations. To illustrate intrinsic distinction, consider the Bunge Euler angle [124] convention—three successive rotations (ϕ_1, Φ, ϕ_2) transform the sample coordinate system into the crystal coordinate system. The orientation of a crystal is expressed as a sequence of these three rotations:

(i) a rotation by ϕ_1 about the z-axis,

(ii) a rotation by Φ about the new x'-axis, and

(iii) a rotation by ϕ_2 about the new z''-axis.

This description is intrinsic, since the axes evolve after each rotation. As illustrated in Figure 3.4, the same orientation can equally be described in extrinsic terms, i.e., as successive rotations about the original, fixed z, x, and z-axes:

(i) a rotation by ϕ_2 about the z-axis,

(ii) a rotation by Φ about the fixed initial x-axis, and

(iii) a rotation by ϕ_1 about the fixed initial z-axis.

Ambiguity arises when rotation sequences are specified without stating which convention is employed. A notation such as "z–x–z rotation" may thus be interpreted differently depending on whether intrinsic or extrinsic axes are assumed. To avoid misinterpretation, it is essential to explicitly state the chosen convention when describing chained rotations in crystallographic analysis.

3.2.3. Matrix Implementation of Pre- and Post-Multiplication

The choice of convention directly affects how rotation matrices are combined. Depending on whether the transformations are interpreted in an intrinsic or extrinsic manner, the corresponding matrices must be applied through pre- or post-multiplication. This subsection outlines the resulting implementation rules and their implications for chained rotations.

The distinction between intrinsic and extrinsic sequences directly affects how rotations are represented in matrix form. Since matrix multiplication is not commutative (which means in most cases, for two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{AB} \neq \mathbf{BA}$), the order in which transformations are applied must be unambiguous.

In the case of post-multiplication, the transformation is written as

$$\boldsymbol{v}' = \mathbf{R}_z(\phi_1)\,\mathbf{R}_x(\Phi)\,\mathbf{R}_z(\phi_2)\,\boldsymbol{v} \tag{3.3}$$

where the vector $\boldsymbol{v}$ is successively rotated by right-multiplying with the matrices in the given order. This convention is commonly used for active rotations.

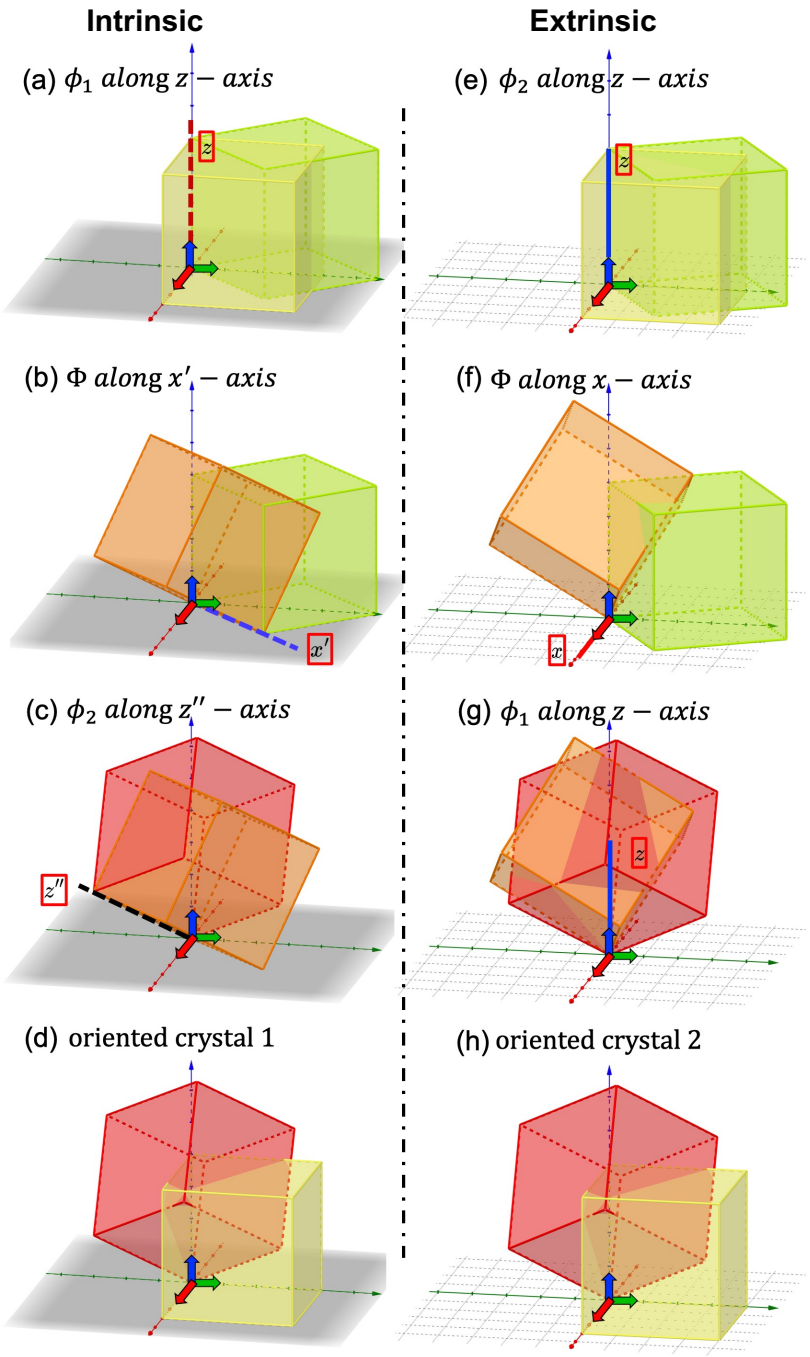

Figure 3.4. Schematic comparison of intrinsic and extrinsic Euler rotations. Panels (a–d) illustrate the intrinsic rotation sequence (45°, 60°, 30°), while panels (e–h) depict the extrinsic rotation sequence (30°, 60°, 45°). In both cases, the three Euler angles remain $\phi_1 = 45°$, $\Phi = 60°$, and $\phi_2 = 30°$. For comprehensive visualization, all rotation operations are available as interactive 3D models online at intrinsic rotations and extrinsic rotations.

By contrast, in pre-multiplication the same orientation is represented as

$$\boldsymbol{v}' = \boldsymbol{v}\,\mathbf{R}_z(\phi_2)\,\mathbf{R}_x(\Phi)\,\mathbf{R}_z(\phi_1), \tag{3.4}$$

so that the matrices appear in reversed order. This convention is often associated with the passive viewpoint.

Confusion frequently arises because the verbal order of rotations (e.g., "rotate z, then x, then z") may not match the algebraic order of matrices in the product. The key is that intrinsic sequences typically correspond to post-multiplication, while extrinsic sequences correspond to pre-multiplication. As long as the chosen convention is explicitly stated, the two approaches are mathematically consistent and yield equivalent results. A detailed discussion and the corresponding Python implementation are also available in the online supplementary materials provided by the Institute of Materials Data Science and Informatics (IAS-9).

3.3. 3D Reconstruction

3.3.1. Uniform Scaling

In TEM imaging, projection occurs inherently during the acquisition process, producing 2D pixel-based images. For subsequent quantitative analysis, the images must first be calibrated using the scale bar to convert pixel distances into physical units. Only after this scaling can the projected dislocation geometry be consistently interpreted and used for 3D reconstruction.

In projective geometry, commonly employed in computer graphics, points are expressed in homogeneous coordinates. In this representation, a 3D point is written as a four-component vector $[x, y, z, 1]^\top$, and transformations are therefore expressed as 4×4 matrices to operate on these 4D homogeneous vectors. Accordingly, a uniform scaling by a factor s can be represented by the transformation matrix

$$\mathbf{S}(s) = \begin{bmatrix} s & 0 & 0 & 0 \\ 0 & s & 0 & 0 \\ 0 & 0 & s & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Case 1: scaling before projection. Let $\mathbf{M}$ denote the projection matrix onto the observation plane. If $\mathbf{S}(s)$ is first applied to a homogeneous point $\boldsymbol{P} = (\boldsymbol{P}_x, \boldsymbol{P}_y, \boldsymbol{P}_z, 1)^\top$, this yields

$$\boldsymbol{P}' = \mathbf{S}(s)\boldsymbol{P} = (s\boldsymbol{P}_x,\, s\boldsymbol{P}_y,\, s\boldsymbol{P}_z,\, 1)^\top,$$

the projected point is then obtained as

$$\boldsymbol{P}_1 = \mathbf{M}\boldsymbol{P}' = \mathbf{M}\mathbf{S}(s)\boldsymbol{P}.$$

Case 2: projection before scaling. Alternatively, if projection is applied first, we compute

$$\boldsymbol{P}'' = \mathbf{M}\boldsymbol{P},$$

followed by uniform scaling,

$$\boldsymbol{P}_2 = \mathbf{S}(s)\boldsymbol{P}'' = \mathbf{S}(s)\mathbf{M}\boldsymbol{P}.$$

The resulting coordinate $\boldsymbol{P}_2$ differs from the result $\boldsymbol{P}_1$ obtained in Case 1.

The discrepancy between $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ demonstrates that scaling and projection do not commute. From a physical perspective, scaling should correspond to rescaling real spatial distances rather than pixel distances after projection. Therefore, the correct order of operations is to apply scaling first and then perform the projection. This ensures that the final coordinates in the glide-plane basis reflect true physical dimensions. Although trivial in computer graphics, this example demonstrates why the order of operations must be defined explicitly before introducing chained rotations in crystallography.

3.3.2. Affine Projection

The 3D information available in the reconstructed tomography allows simple identification of the slip plane and line direction of the dislocations, allowing more direct characterization of the local stress state as well as the true spatial distribution of the defects.

In the experiment, the CCS and SCS are linked through Euler angles which describe a sequence of rotations for one coordinate system to match/overlap with the other. Our TEM experimental data was processed with the open source software *pycotem* [125] where right-handed coordinate system is used; active rotation matrix is used; Bunge convention is used and intrinsic rotation convention is used. That is to say, for a given Euler angle (ϕ_1,Φ,ϕ_2), the corresponding rotation matrix reads

$$\mathbf{R}(z''\,|\,\phi_2,\ x'\,|\,\Phi,\ z\,|\,\phi_1) = \mathbf{R}^{\mathrm{a}}_z(\phi_1)\ \mathbf{R}^{\mathrm{a}}_x(\Phi)\ \mathbf{R}^{\mathrm{a}}_z(\phi_2). \tag{3.5}$$

In the above equation, on the left-hand side $\mathbf{R}$ is an operator which executes from the right to the left, while on the right-hand side is the chain rotation matrix with $\mathbf{R}^{\mathrm{a}}_z$ and $\mathbf{R}^{\mathrm{a}}_x$ being the active rotation matrix about z- and x-axis respectively.

Having clarified the Euler angle conventions, we are able to express/calculate dislocation related information such as Burgers vector and slip plane normal in the SCS, and finally in WCS after further tilting. Having all information in WCS, reconstructing dislocation microstructure is essentially projecting dislocation lines on the viewing screen back to the slip plane, as shown in Figure 3.5. For a given point $\boldsymbol{P}_i$ on the slip plane, its projection on the viewing screen is denoted by $\boldsymbol{V}_i$, such that

$$\boldsymbol{P}_i = \boldsymbol{V}_i + \boldsymbol{R}d_i. \tag{3.6}$$

Here, $\boldsymbol{R}$ is unit vector along the reverse electron beam direction/global z_{w}-axis, i.e., [001]. On the slip plane, we have

$$(\boldsymbol{P}_i - \boldsymbol{P}_0)\cdot\boldsymbol{n}_{\mathrm{w}} = 0, \tag{3.7}$$

where $\boldsymbol{P}_0$ is the position vector of an arbitrary point on the slip plane, and $\boldsymbol{n}_\mathrm{w}$ is the slip plane normal under WCS. The term $(\boldsymbol{P}_i - \boldsymbol{P}_0)$ thus represents a vector lying within the slip plane. Combining the above two equations, we obtain

$$d_i = \frac{\boldsymbol{n}_\mathrm{w} \cdot (\boldsymbol{P}_0 - \boldsymbol{V}_i)}{\boldsymbol{n}_\mathrm{w} \cdot \boldsymbol{R}}. \tag{3.8}$$

For the convenience of the calculation, we let $\boldsymbol{P}_0$ coincide with $\boldsymbol{V}_0$ as the origin on the viewing screen. After solving d_i, any point $\boldsymbol{V}_i$, which are dislocation polygon's vertices, can be projected back to the slip plane.

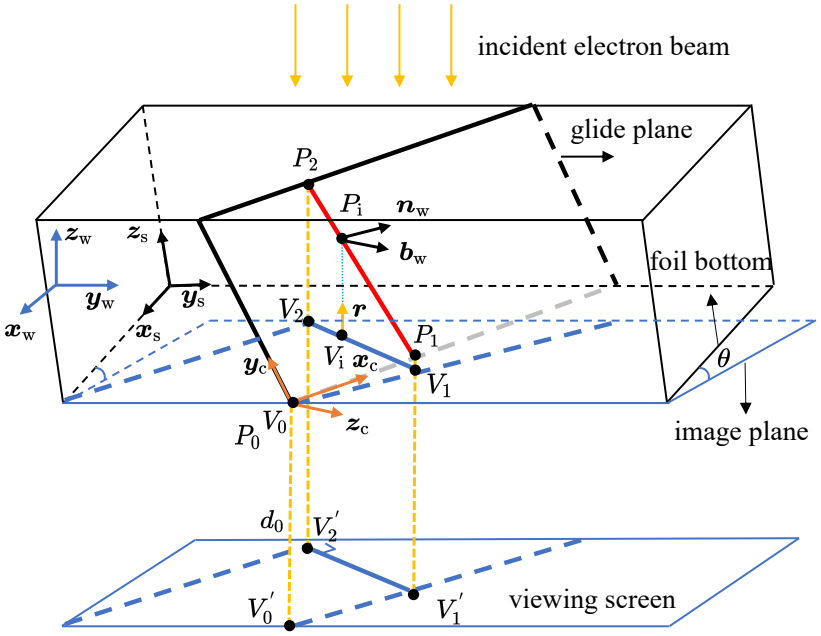

Figure 3.5. The planar and linear feature in a foil projected on the image plane. The plane normal $\boldsymbol{n}_\mathrm{w}$ is determined by the reversed beam direction $\boldsymbol{R}$, projected line $\boldsymbol{l}$ or trace direction $\boldsymbol{P}_0\boldsymbol{P}_1$, projected trace direction $\boldsymbol{V}_0\boldsymbol{V}_1$ and projected plane width V_1V_2, at different tilts θ along y_w-axis.

For all dislocation coordinates $\boldsymbol{P}_i$ in 3D, we further project them onto the slip plane, where $\boldsymbol{P}_0$ is taken as the origin, the Burgers vector $\boldsymbol{b}_\mathrm{w}$ is chosen as the local x-axis, and slip plane normal $\boldsymbol{n}_\mathrm{w}$ as local z-axis. The resulting dislocation coordinates on the slip plane are denoted by $\boldsymbol{P}_i^\mathrm{sp}$,

$$\boldsymbol{P}_i^\mathrm{sp} = \Big[\boldsymbol{b}_\mathrm{w} \cdot (\boldsymbol{P}_i - \boldsymbol{P}_0),\ (\boldsymbol{n}_\mathrm{w} \times \boldsymbol{b}_\mathrm{w}) \cdot (\boldsymbol{P}_i - \boldsymbol{P}_0),\ 0\Big]. \tag{3.9}$$

3.3.3. Validation of the Formula

For intuitive understanding and interactive exploration of spatial geometry, a simplified 3D toy model was constructed in *GeoGebra* [126] to simulate the orientation of a cubic crystal in the world coordinate system (WCS) and to visualize slip plane configurations

under TEM-holder tilting conditions. The model further yields a 2D projection of slip traces on the WCS xy-plane, corresponding to the TEM screen image. The modeling steps are as follows:

(1) A cube with an edge length of 4 units (thickness $t = 4$) is defined as the crystal (in Figure 3.6a, light yellow). The vertices of the base plane are labeled A, B, C, and D (counterclockwise from the inner left corner), and those of the top face are E, F, G, and H. A slip plane $ABKL$ is introduced, where K, L, M, and N denote the midpoints of edges FG, EH, BC, and AD, respectively. The Miller index of this slip plane is identified as $(1, 2, \bar{1})$, and the two slip traces are KL and MN. The dislocation line l is defined parallel to edge KM. In this particular case, the angle α, defined as that between the *dislocation line* and the *image plane*, coincides with the angle between the *slip plane* and the *image plane*.

(2) The crystal orientation relative to the sample coordinate system (SCS) is defined by Bunge Euler angles

$$\phi_1 = -60^\circ, \quad \Phi = 30^\circ, \quad \phi_2 = 45^\circ.$$

After applying the intrinsic z–x'–z'' rotations, the crystal (colored light red) in the SCS is shown in Figure 3.6b.

(3) Initially, the SCS is assumed to coincide with the WCS. The specimen is then tilted by $\theta = -45^\circ$ (counterclockwise, following the right-hand rule) about the y_w-axis, representing a single-tilt operation of the TEM holder. The resulting crystal (light blue cubic) is shown in Figure 3.6c, while Figure 3.6d and Figure 3.6e displays the 2D projection of the slip plane trace on the WCS xy-plane (image plane) in 3D view, and 2D view, respectively.

Based on the geometrical relations (as simplified in Figure 3.7). The specimen thickness t is given by

$$t = PS = PQ \cdot \sin\beta, \qquad PQ = \frac{QT}{\cos\alpha} = \frac{L_{\text{proj}}}{\cos\alpha}.$$

Here, β is the angle between the *dislocation line* and the *crystal base plane*, α is the angle between the *dislocation line* and the *image plane* (not between the slip plane and the image plane), and L_{proj} is the perpendicular spacing between the two slip traces on the world base plane (i.e., the image plane). Combining these relations yields

$$t = L_{\text{proj}} \cdot \frac{\sin\beta}{\cos\alpha}. \tag{3.10}$$

Alternatively, the thickness can also be calculated in crystal coordinate system (CCS) by

$$t = d_{\min} \cdot \tan\beta, \tag{3.11}$$

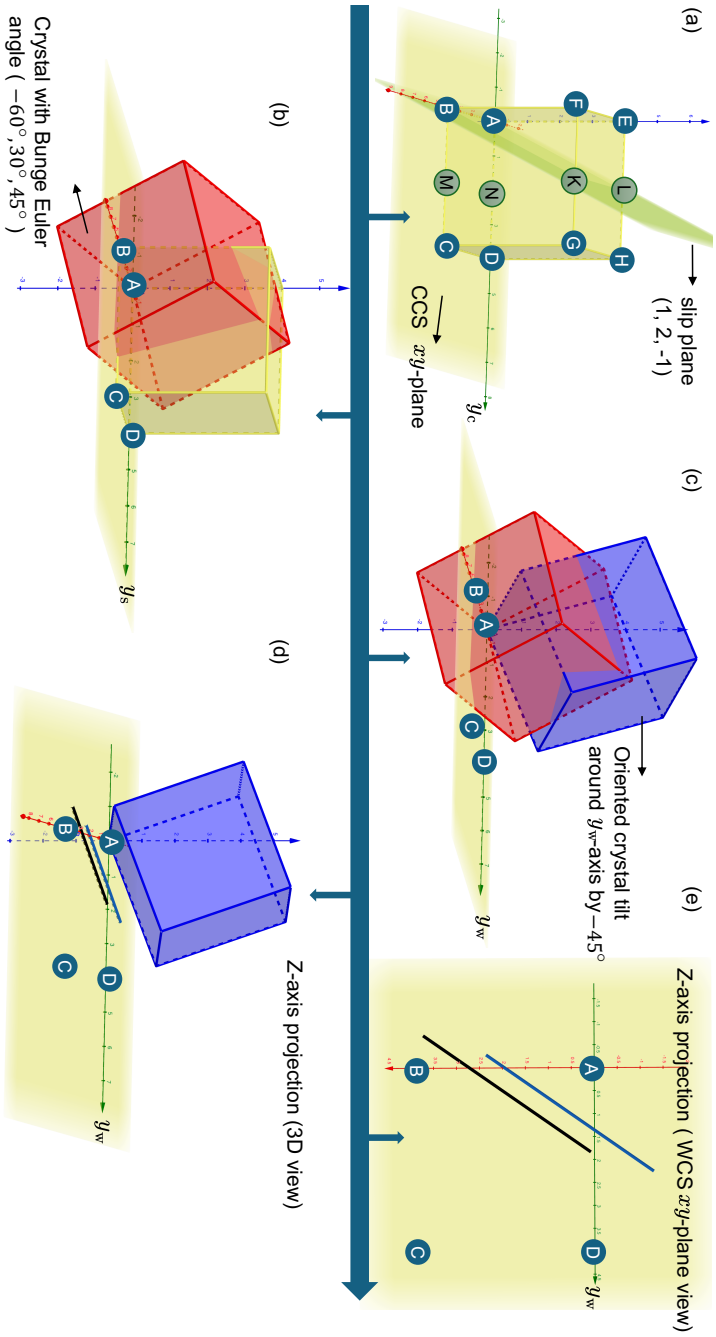

Figure 3.6. Geometrical description of the crystal and its location in different coordinate systems. (a) Cubic crystal with slip plane $ABKL$ - Miller index: (1 2 $\bar{1}$). (b) Orientation in the SCS with Euler angles $\phi_1 = -60°$, $\Phi = 30°$, $\phi_2 = 45°$. (c) Orientation in the WCS after tilt about y_w by $\theta = -45°$. (d) Projection of the slip plane trace onto the WCS xy-plane (3D view). (e) Projection of the slip plane trace onto the WCS xy-plane (2D view). An interactive version of this figure is available online at GeoGebra.

where $d_{\min}$ is the minimum (perpendicular) distance between the two slip traces on the crystal base plane. This makes physical sense, and the relation between $d_{\min}$ and L_{proj} can be obtained as:

$$\begin{aligned} \cos\beta &= d_{\min}/l \\ &\Rightarrow \quad L_{\text{proj}} = d_{\min} \cdot \frac{\cos\alpha}{\cos\beta}. \\ \cos\alpha &= L_{\text{proj}}/l \end{aligned} \tag{3.12}$$

Here, l denotes the length of the dislocation line.

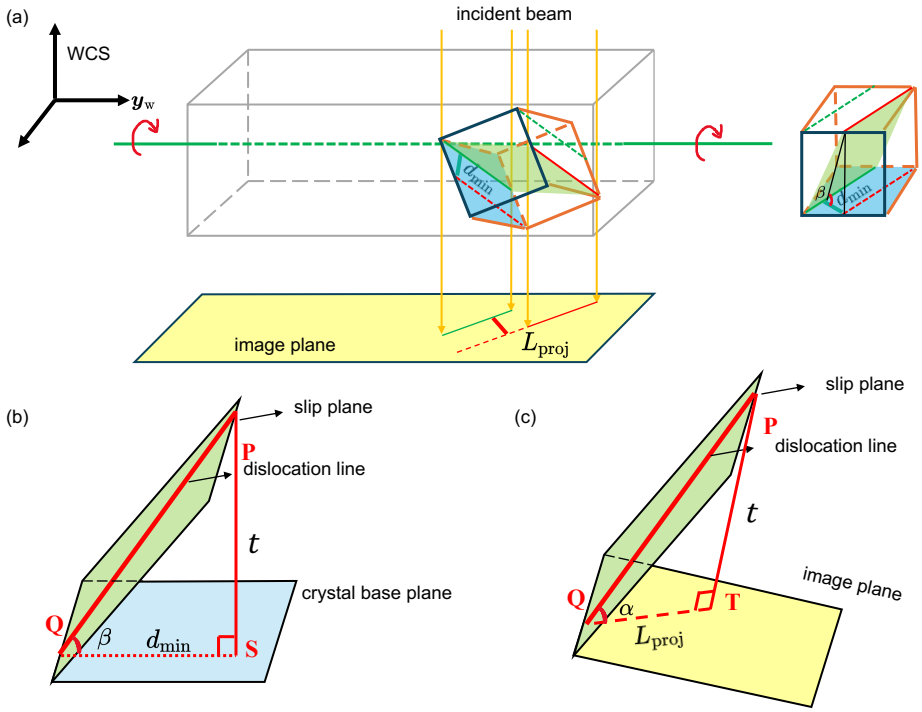

Figure 3.7. Geometrical setup and thickness estimation in different coordinate systems. (a) Simplified geometry of the tilted crystal in WCS. (b) Thickness estimation in CCS. (c) Thickness estimation in WCS.

Both approaches give consistent values, with $t_1 = 3.958$ and $t_2 = 3.999$, the minor discrepancy (in initial setting $t = 4$) is due to rounding in *GeoGebra*. The definitions and numerical values of key parameters are summarized in Table 3.1. This consistency validates the correctness of both the geometrical and the projection-based constructions, thereby supporting their use in the subsequent analysis of TEM dislocation reconstructions.

3.3.4. Discussion of Geometrical Formulations in Previous Methods

The present geometrical treatment is not intended to deliver an exact solution to TEM thickness determination. In practice, the dynamical and non-linear electron effect,

Table 3.1. Definition and numerical values of symbols in the toy model

Symbol	Definition	Value
(ϕ_1, Φ, ϕ_2)	Bunge Euler angles	$(-60°, 30°, 45°)$
$\boldsymbol{n}_c$	Slip-plane normal vector in the CCS	$[1, 2, \bar{1}]$
θ	Tilt angle (right-hand convention)	$-45°$
L_{proj}	Minimum distance between the two slip traces on the image plane	3.95
α	Angle between dislocation line and image plane	84.31°
β	Angle between dislocation line and crystal base plane	114.09°
d_{min}	Minimum distance between the two slip traces on the crystal base plane	1.79

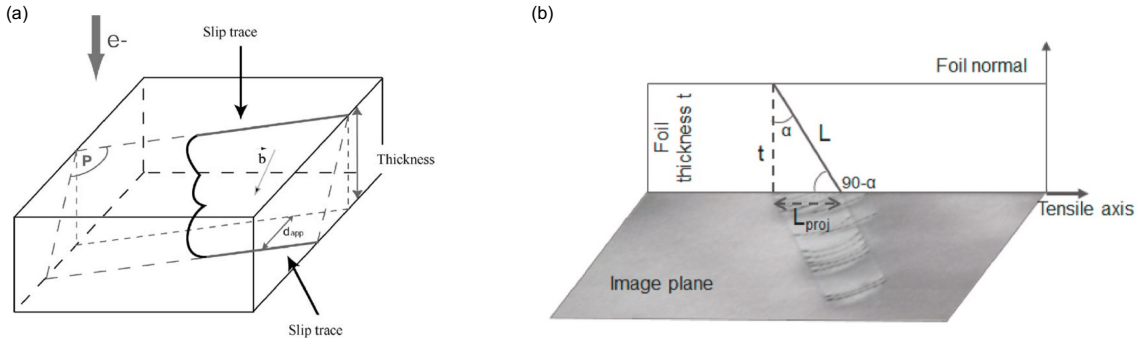

Figure 3.8. Schematic representations of specimen-thickness estimation methods reported in the literature, with the associated expressions as stated in the original constructions. (a) $t = L_{\text{proj}} \frac{\sin \alpha'}{\cos(\alpha' - \theta)}$ (after Delmas et al. [129]). (b) $t = \frac{L_{\text{proj}}}{\tan(90° - \alpha'')}$ (after Oliveros et al. [130]).

multiple scattering, and even the preparation of the sample influence image formation beyond the scope of a simplified geometric framework [127, 128]. Nevertheless, from a materials-science perspective, it is valuable to examine the internal consistency of analytical geometric constructions that are widely used for approximate thickness estimation.

In this context, we note that the simplified formulas reported in previous works [129, 130] are practically useful but contain geometrical ambiguities in their derivation. This observation does not invalidate their use as approximations; rather, it highlights the risk of misinterpretation if the underlying angle definitions and reference frames are not stated explicitly. Below we clarify the assumptions and limits of applicability.

Remarks on the formulation by Delmas et al. [129]. The expression is often written in the form

$$t \;=\; L_{\mathrm{proj}}\,\frac{\sin\alpha'}{\cos(\alpha'-\theta)},$$

which implicitly mixes two angles defined in *different* reference relationships. In Delmas' construction (with variables sharing the same definitions rewritten using the notation adopted in this work), α' is taken as the angle between the slip plane and the *sample* surface (SCS), whereas in our framework α is defined as the angle between dislocation line l with respect to the *crystal* base plane (CCS). Unless the angles α' and θ are coplanar with a shared edge, the subtraction $(\alpha'-\theta)$ has no clear geometric meaning. Consequently, the formula can be used consistently only in special cases (e.g., $\theta = 0°$ with CCS aligned to SCS and with an explicitly stated definition of α). In the general case, the tilt θ enters independently through the projection geometry and should not be combined with α by a direct subtraction unless the coplanarity and edge-sharing conditions are first established.

Remarks on the formulation by Oliveros et al. [130]. A projection-based expression of the form

$$t \;=\; \frac{L_{\mathrm{proj}}}{\tan(90°-\alpha'')},$$

(or other forms involving a distance denoted by $d_{\min}$) relies on extra information that is not, in general, directly accessible from projection onto the image plane. In particular, the angle between the *slip plane* and the *image plane* is almost never equal to the angle between the *dislocation line in the slip plane* and the *image plane*, the latter is that matters when constructing the thickness formula. Moreover, the explicit dependence on the specimen tilt θ is absent in this form, which limits its applicability when the viewing direction deviates from the reference base plane.

Practical implication. For routine, single y_{w}-axis tilt TEM imaging, a consistent geometric treatment separates (i) the angle between *dislocation line* and the *image plane* (ii) the specimen tilt defined in the laboratory frame, leading to a thickness expression of the form

$$t \;=\; L_{\mathrm{proj}}\cdot\frac{\sin\alpha}{\cos\theta},$$

where the angle definitions and reference frames are stated explicitly. When angle definitions differ from those above (e.g., α referenced to the angle between two planes rather than between line and the crystal base), the relations must be re-derived in the corresponding frame. While multiple single-tilt acquisitions at different orientations can, in principle, provide a more robust basis for validation, the present section is limited to the geometrical clarification. A separate, independent thickness validation using electron energy loss spectroscopy (EELS) is given in Chapter 6 (TEM pile-up).

3.4. Summary

This chapter presented a consistent geometrical framework for the three-dimensional reconstruction of dislocations in TEM. By defining crystallite orientations and reference coordinate systems, and by clarifying chained rotation conventions, we established unambiguous tools for handling orientation transformations. The framework was demonstrated through simplified projection geometry and a toy example, and was further used to clarify ambiguities in previously reported thickness formulas. A direct comparison with existing analytical methods reveals that several commonly used expressions rely on implicit assumptions about reference frames and rotation conventions, which can lead to geometrical inconsistencies if not stated explicitly. These results provide a transparent basis for interpreting TEM projections and set the stage for experimental validation in the following chapter.

4. Computational Characterization of Dislocation Microstructure

To advance the understanding of the motion of collective dislocations, it is essential to complement direct experimental observations with computational approaches. *In situ* TEM provides valuable insights into the trajectories of moving dislocations, yet the complexity of their interactions, such as pinning and depinning events, cannot be fully resolved by imaging alone. A data-driven characterization of such processes requires a rigorous theoretical and numerical foundation that bridges discrete dislocation behavior with continuum-level descriptions. In this context, discrete dislocation dynamics (DDD) offers a means to explicitly track the evolution of dislocation networks, while continuum dislocation dynamics (CDD) provides a coarse-grained framework for field-based representations. The discrete-to-continuum (D2C) methodology further establishes a systematic link between these two scales. Building upon these foundations, this chapter validates computational tools for characterizing dislocation microstructures from spatio–temporal perspectives.

4.1. Dislocation Dynamics

Dislocation dynamics provides a theoretical and computational foundation for investigating the collective behavior of dislocation ensembles across length and time scales. Two complementary approaches have been developed: DDD, which explicitly resolves the motion of individual dislocations, and CDD, which reformulates the same physics in terms of coarse-grained field variables. Together, these approaches enable multiscale analyses that connect atomistic mechanisms with continuum-scale material responses.

4.1.1. Discrete Dislocation Dynamics

The conceptual foundations of dislocation dynamics trace back to the 1930s, when the independent works of Taylor, Orowan, and Polanyi established the dislocation as the carrier of plastic deformation. In the following decades, theoretical studies of dislocation interactions laid the groundwork for numerical treatments [131, 40, 41, 42]. With the advent of modern computing in the 1980s–1990s, these ideas evolved into discrete dislocation dynamics (DDD), in which dislocations are represented as line segments in an elastic medium and their motion is governed by the Peach–Koehler force [50]. Subsequent developments, contributed by Bulatov [132], Cai [133], Ghoniem [134], Kubin [135] and many others [136, 137, 138] led to robust node-segment algorithms capable

of tracking complex dislocation networks. Modern frameworks such as *ParaDiS* [139] and *MicroMegas* [140] further incorporate fast multipole methods [141] for long-range stress evaluation and non-singular kernel formulations, allowing simulations of systems containing millions of dislocation segments. Today, DDD is still regarded as a central tool for linking microscale dislocation mechanisms to macroscopic plasticity [142].

In DDD, the fundamental equation of motion assumes an overdamped regime, where the velocity of each dislocation segment is proportional to the Peach–Koehler force acting on it,

$$\boldsymbol{v} = \mathbf{M} \cdot \boldsymbol{F}, \qquad \boldsymbol{F} = (\boldsymbol{\sigma} \cdot \boldsymbol{b}) \times \boldsymbol{\xi}. \tag{4.1}$$

Here, $\boldsymbol{v}$ is the velocity of the dislocation segment, $\mathbf{M}$ the mobility tensor, $\boldsymbol{F}$ the Peach–Koehler force, $\boldsymbol{\sigma}$ the local stress tensor, $\boldsymbol{b}$ the Burgers vector, and $\boldsymbol{\xi}$ the unit tangent vector along the dislocation line, which is equivalent to $\boldsymbol{l}^{\mathrm{d}}$ introduced in Equation (4.2). The different notation is adopted here to follow the conventional form of the Peach–Koehler force expression in the DDD literature. This overdamped formulation provides the basis for modeling topological reactions such as annihilation, junction formation, cross-slip, and climb [143].

Despite its power, DDD faces intrinsic limitations. The computational cost grows rapidly with system size, restricting simulations to micrometer scales and microsecond times. Furthermore, boundary conditions and mobility laws often require empirical input, reducing predictive capability. As a result, while DDD excels at capturing detailed dislocation interactions and mesoscale plasticity phenomena, it cannot be straightforwardly extended to macroscopic material volumes.

4.1.2. Continuous Dislocation Dynamics

To overcome the scale limitations of DDD, continuum formulations of dislocation dynamics have been pursued for decades. The foundation lies in Kröner's and Nye's tensorial descriptions of lattice curvature and geometrically necessary dislocations (GNDs) [144, 145, 146, 147]. Later, El-Azab [148] introduced a statistical framework for describing dislocation ensembles, in which the state of many dislocations within a representative volume is characterized by a density distribution function rather than by tracking individual lines. This formalism enabled the derivation of conservation-type, or transport equations, governing the spatio-temporal evolution of dislocation densities. Building upon these concepts, Hochrainer and co-workers [149, 150, 151, 118, 152, 153] developed higher-order CDD, where dislocation density is treated as a tensor-valued field, and additional state variables (such as line orientation distributions) are introduced to capture microstructural anisotropy.

In CDD, the discrete network of dislocation lines is replaced by continuous density fields whose evolution is governed by transport equations. This field-based description enables coarse-grained analyses of dislocation-mediated plasticity from mesoscopic to macroscopic scales and facilitates integration with continuum plasticity models.

Despite these advantages, CDD has inherent limitations. Closure assumptions are required to truncate the hierarchy of evolution equations, which introduces uncer-

tainties. Moreover, the parameterization of mobility laws and interaction terms still relies heavily on input from lower-scale simulations or experiments. Consequently, while CDD provides a powerful framework for bridging to macroscopic constitutive models, its predictive capability ultimately depends on the accuracy of these underlying approximations.

4.2. Discrete-to-Continuum Dislocation Dynamics

To connect discrete dislocation configurations with continuum-level descriptions, it is necessary to represent line geometries in terms of field variables. A systematic formalism for this purpose was first established in the statistical dislocation theories of Groma and co-workers [154, 155, 156, 157, 158, 159], and was subsequently advanced and generalized within the discrete-to-continuum (D2C) framework by Sandfeld and co-workers [4, 160].

In the present work, we make use of only part of the D2C framework, specifically the definitions of coarse-grained dislocation density and curvature density, while other aspects of the full theory are beyond the scope of this study. In addition, following the same coarse-graining strategy, we introduce the coarse-grained velocity field, which, although not explicitly included in the original D2C formulation, serves as a useful quantity for analyzing the kinematics of dislocation networks.

4.2.1. Motivation: Bridging Dislocation Dynamics

DDD simulations provide detailed information on individual dislocation segments, but the resulting data are difficult to analyze statistically and cannot be directly compared with continuum models or experiments. In contrast, CDD offers transport equations for dislocation density fields, but these equations require closure assumptions and calibrated input that are not directly accessible from lower scale models. Without a systematic mapping, the two approaches remain disconnected. In the literature, D2C has been demonstrated to enable three key advances:

- the extraction of density, curvature, and velocity fields from discrete dislocation networks [4, 160, 161],
- the validation and calibration of continuum models against discrete simulations [4, 119, 162, 163], and
- the establishment of a framework that connects numerical simulations with experimental observations, such as *in situ* TEM [120, 164].

Thus, D2C provides a crucial link between discrete and continuum scales, enabling detailed microscale information to inform mesoscale and macroscale models.

4.2.2. Parametric Discrete Dislocation Line Representation

A dislocation line can be represented as an oriented curve $\boldsymbol{c}(\ell)$ parameterized by its arc length ℓ, as illustrated in Figure 4.1. The local tangent and curvature vectors are given by

$$\boldsymbol{l}^{\mathrm{d}}(\ell) = \frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}\ell}, \qquad \boldsymbol{k}^{\mathrm{d}}(\ell) = \frac{\mathrm{d}^2\boldsymbol{c}}{\mathrm{d}\ell^2} \tag{4.2}$$

Here, the superscript "d" indicates discrete quantities, and it is omitted when referring to their continuum counterparts or when the context is clear.

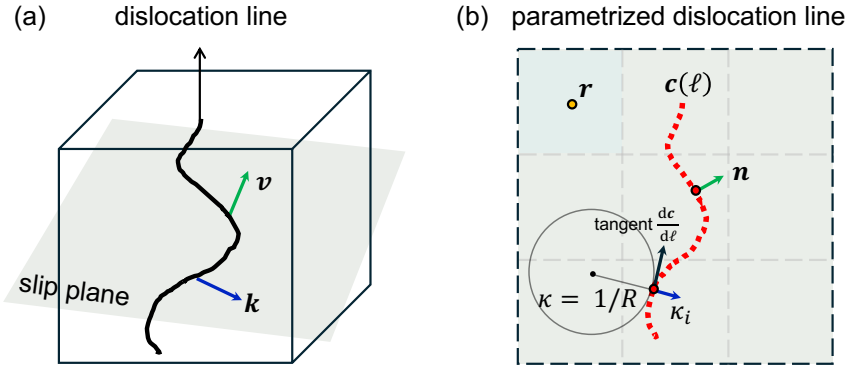

Figure 4.1. Schematic illustration of a parametrized discrete dislocation line $\boldsymbol{c}(\ell)$ and its geometric descriptors. (a) The dislocation line together with the curvature vector $\boldsymbol{k}^{\mathrm{d}} = \mathrm{d}^2\boldsymbol{c}/\mathrm{d}\ell^2$ and the velocity vector $\boldsymbol{v}$, characterizes the dislocation. (b) The local tangent vector $\boldsymbol{l}^{\mathrm{d}} = \mathrm{d}\boldsymbol{c}/\mathrm{d}\ell$ is shown together with the curvature $\kappa = \|\boldsymbol{k}^{\mathrm{d}}\| = 1/R$, the slip-plane normal $\boldsymbol{n}$, and a spatial point $\boldsymbol{r}$ for reference.

Since the arc-length parameter ℓ is only available after the curve has been constructed, it is often convenient to prescribe a curve through an auxiliary parameter $u \in [0, 1]$. In this case, the tangent and curvature can be derived via u,

$$\boldsymbol{l}^{\mathrm{d}}(u) = \frac{1}{J}\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}u}, \qquad \boldsymbol{k}^{\mathrm{d}}(u) = \frac{1}{J^2}\frac{\mathrm{d}^2\boldsymbol{c}}{\mathrm{d}u^2} - \frac{1}{J^3}\frac{\mathrm{d}J}{\mathrm{d}u}\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}u}. \tag{4.3}$$

Here, the scalar J and its derivative are defined as

$$J(u) = \frac{\mathrm{d}\ell}{\mathrm{d}u} = \left\|\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}u}\right\|, \qquad \frac{\mathrm{d}J}{\mathrm{d}u} = \frac{1}{J}\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}u}\cdot\frac{\mathrm{d}^2\boldsymbol{c}}{\mathrm{d}u^2}. \tag{4.4}$$

4.2.3. Averaging Dislocation Fields

As an auxiliary functional for averaging geometrical properties of discrete dislocations, we introduce the Dirac delta operator on a dislocation line $\boldsymbol{c}$ as

$$\delta_{\boldsymbol{c}}([\circledast], \boldsymbol{r}) = \int_0^{L_{\boldsymbol{c}}} \delta(\boldsymbol{c}(\ell) - \boldsymbol{r}) \circledast \mathrm{d}\ell. \tag{4.5}$$

Here, the symbol $\circledast$ denotes an arbitrary line property (scalar, vector, or tensor function), and $\boldsymbol{r}$ is a spatial point.

Using this definition, the scalar dislocation density $\rho_{\boldsymbol{c}}$ and the curvature density vector $\boldsymbol{q}_{\boldsymbol{c}}$ of a single line $\boldsymbol{c}$ are expressed as

$$\rho_{\boldsymbol{c}}(\boldsymbol{r}) = \int_0^{L_{\boldsymbol{c}}} \delta(\boldsymbol{c}(\ell) - \boldsymbol{r})\,\mathrm{d}\ell. \tag{4.6}$$

$$\boldsymbol{q}_{\boldsymbol{c}}(\boldsymbol{r}) = \int_0^{L_{\boldsymbol{c}}} \delta(\boldsymbol{c}(\ell) - \boldsymbol{r})\frac{\mathrm{d}^2\boldsymbol{c}}{\mathrm{d}\ell^2}\,\mathrm{d}\ell. \tag{4.7}$$

The scalar, signed curvature density $q_{\boldsymbol{c}}$ can be obtained by projecting the curvature density vector $\boldsymbol{q}_{\boldsymbol{c}}$ onto the outward slip-plane normal,

$$q_{\boldsymbol{c}}(\boldsymbol{r}) = \int_0^{L_{\boldsymbol{c}}} \delta(\boldsymbol{c}(\ell) - \boldsymbol{r})\frac{\mathrm{d}^2\boldsymbol{c}}{\mathrm{d}\ell^2} \cdot \left(\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}\ell} \times \boldsymbol{n}\right)\mathrm{d}\ell \tag{4.8}$$

where $\boldsymbol{n}$ is the unit normal to the slip plane.

These discrete measures are well suited for spatial averaging over multiple lines. For this purpose, we define an averaging operator over a representative volume $V_{\boldsymbol{r}}$ of size V, centered around position $\boldsymbol{r}$,

$$\langle \circledast \rangle_{V,\boldsymbol{r}} := \frac{1}{V}\int_{V_{\boldsymbol{r}}} \circledast\, \mathrm{d}^3 r. \tag{4.9}$$

By applying this coarse-graining operator, the coarse-grained dislocation density $\langle \rho \rangle (\boldsymbol{r})$ and curvature density $\langle q \rangle (\boldsymbol{r})$ fields for a network of dislocation lines are derived from their single-line counterparts given in Equation (4.6) and Equation (4.8) as

$$\langle \rho \rangle (\boldsymbol{r}) = \frac{1}{V}\sum_{\boldsymbol{c}} \int_{\mathcal{L}_{\boldsymbol{c}}^{V,\boldsymbol{r}}} \mathrm{d}\ell, \tag{4.10}$$

$$\langle q \rangle (\boldsymbol{r}) = \frac{1}{V}\sum_{\boldsymbol{c}} \int_{\mathcal{L}_{\boldsymbol{c}}^{V,\boldsymbol{r}}} \frac{\mathrm{d}^2\boldsymbol{c}}{\mathrm{d}\ell^2} \cdot \left(\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}\ell} \times \boldsymbol{n}\right)\mathrm{d}\ell, \tag{4.11}$$

where $\mathcal{L}_{\boldsymbol{c}}^{V,\boldsymbol{r}} \subset \boldsymbol{c}$ denotes the portion of line $\boldsymbol{c}$ contained inside the averaging volume V centered at $\boldsymbol{r}$.

The same coarse-graining framework can also be applied to other line-associated quantities. In particular, the coarse-grained glide velocity field $\langle v \rangle (\boldsymbol{r})$ can be expressed as

$$\langle v \rangle (\boldsymbol{r}) = \frac{1}{V}\sum_{\boldsymbol{c}} \int_{\mathcal{L}_{\boldsymbol{c}}^{V,\boldsymbol{r}}} \boldsymbol{v} \cdot \left(\frac{\mathrm{d}\boldsymbol{c}}{\mathrm{d}\ell} \times \boldsymbol{n}\right)\mathrm{d}\ell, \tag{4.12}$$

where $\boldsymbol{v}$ is the local line velocity and $\boldsymbol{n}$ the unit normal to the slip plane.

4.2.4. Numerical Implementation of the Spatio-Temporal Averaging

For numerical implementation, the domain is discretized into voxels (or pixels, in 2D) with centers $\boldsymbol{r}_i$ located at coordinates (x_i, y_i, z_i). Each voxel defines an averaging subdomain Ω_i as

$$\Omega_i = \left[x_i - \tfrac{1}{2}\Delta x,\, x_i + \tfrac{1}{2}\Delta x\right] \times \left[y_i - \tfrac{1}{2}\Delta y,\, y_i + \tfrac{1}{2}\Delta y\right] \times \left[z_i - \tfrac{1}{2}\Delta z,\, z_i + \tfrac{1}{2}\Delta z\right]. \tag{4.13}$$

Here, the subvolume $V_i = \Delta x\, \Delta y\, \Delta z$ specifies the spatial resolution of the coarse-grained continuum representation.

For each voxel, the coarse-grained dislocation density ρ_i and curvature density q_i can be computed. The average curvature is then given by

$$\kappa_i = \frac{q_i}{\rho_i}. \tag{4.14}$$

For a given frame at time $t = j$, we denote the voxel-averaged curvature and velocity by κ_i^j and v_i^j, respectively. To obtain temporal averages, a density-weighted summation is performed over all frames,

$$\bar{\kappa}_i = \frac{\sum_{j=0}^{n} \rho_i^j \kappa_i^j}{\sum_{j=0}^{n} \rho_i^j}, \qquad \bar{v}_i = \frac{\sum_{j=0}^{n} \rho_i^j v_i^j}{\sum_{j=0}^{n} \rho_i^j}. \tag{4.15}$$

Here, n is the total number of time frames considered.

In summary, this framework establishes a consistent procedure for translating discrete dislocation lines into coarse-grained field variables. By parameterizing line geometry, introducing a Dirac delta operator, and applying spatio–temporal averaging, one can define density, curvature, and velocity fields on a voxelized domain. These fields provide a quantitative basis for comparing discrete trajectories from TEM or DDD simulations with continuum-level descriptions such as CDD, thereby enabling a unified characterization of dislocation microstructures across scales.

4.3. Test Cases

To verify the voxel-based coarse-graining framework introduced in the previous section, we design analytically tractable test cases. These simplified geometries serve as ground truth for curvature and velocity fields, allowing us to systematically evaluate the three fundamental requirements of numerical schemes, namely correctness, consistency, and convergence (the "3Cs"). All validations are carried out on a voxelized (2D pixel) domain, ensuring direct comparability with experimental TEM image data.

4.3.1. Validation of Curvature, Density, and Curvature Density

Definition with pixel normalization. In the discrete-to-field mapping, each curve is discretized into small arc-length segments Δs_i with local curvature κ_i. The contribution of these segments to a pixel centered at (I_x, I_y) with area ΔA is accumulated as

$$\rho(I_x, I_y) = \frac{1}{\Delta A} \sum_i \Delta s_i \, \mathbf{1}_{\{(x_i,y_i)\in \text{pixel}(I_x,I_y)\}},$$

$$q(I_x, I_y) = \frac{1}{\Delta A} \sum_i \kappa_i \, \Delta s_i \, \mathbf{1}_{\{(x_i,y_i)\in \text{pixel}(I_x,I_y)\}},$$

$$\kappa_{\text{cg}}(I_x, I_y) = \frac{q(I_x, I_y)}{\rho(I_x, I_y)} \quad \text{if } \rho > 0.$$

Here, $\mathbf{1}_{\{.\}}$ is the indicator function, which equals to 1 if the segment i falls into the given pixel and 0 otherwise. It serves as the discrete analogue of the Dirac delta function used in the continuous definitions of Equation (4.5), with the role of localizing the contribution of each curve segment to its spatial cell. Thus ρ is the line-length density (total line length per pixel area), q is the curvature density (curvature-weighted line length per pixel area), and κ_{cg} is the coarse-grained signed curvature.

As a benchmark, we consider the analytic S-shaped curve

$$y(x) = 0.6 \sin(\pi x) + 0.25 \sin(2\pi x), \quad x \in [-1, 1],$$

with analytic curvature

$$\kappa(x) = \frac{y''(x)}{[1 + (y'(x))^2]^{3/2}}.$$

The curve is interpolated by $n_{\text{ctrl}} = 25$ control points using a cubic spline and the signed numerical curvature κ_{num} is computed, the results are shown in Figure 4.2.

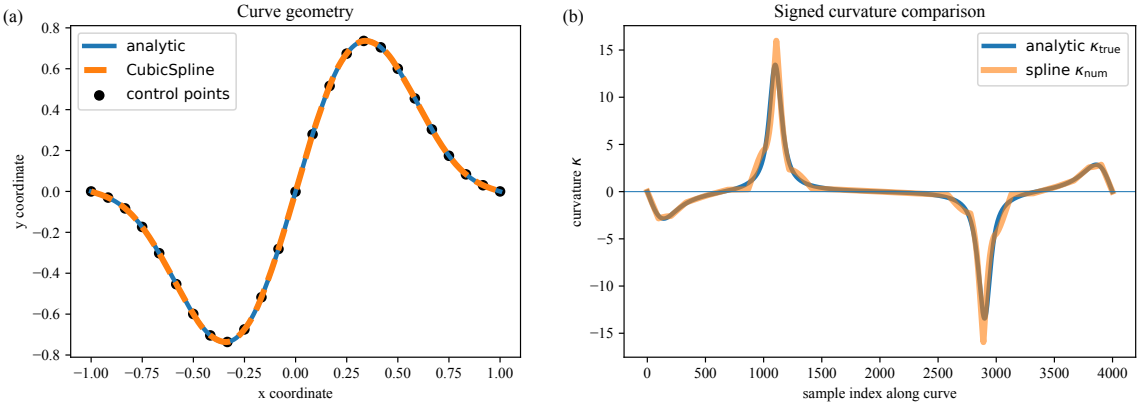

Figure 4.2. Comparison between an analytic curve and its cubic spline interpolation. (a) Analytic curve and interpolated cubic spline. (b) Comparison of the signed curvature.

To assess accuracy, we compute two error norms between κ_{num} and the analytic reference κ_{true}:

- The L^2 norm (root-mean-square error),

$$L^2 = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\kappa_{\mathrm{num},j} - \kappa_{\mathrm{true},j}\right)^2},$$

  which quantifies the average deviation over all N sample points.

- The L^∞ norm (maximum error),

$$L^\infty = \max_{j=1,\dots,N}\left|\kappa_{\mathrm{num},j} - \kappa_{\mathrm{true},j}\right|,$$

  which measures the worst-case local discrepancy.

In our test, the values are

$$L^2 \approx 5.12 \times 10^{-1}, \qquad L^\infty \approx 2.77.$$

These errors are small relative to the dynamic range of the analytic curvature, which varies approximately between −15 and +15 across the domain. The average deviation (L^2 error) thus corresponds to less than 3.5% of the total curvature range. Such error levels are within the expected bounds for a second-derivative quantity computed from a cubic spline based on only 25 control points, and the accuracy can be further improved by increasing the sampling density. Moreover, the numerical curvature reproduces the correct oscillation pattern, sign structure, and the locations of extrema observed in the analytic reference. Together, these observations support the correct implementation of the signed curvature definition in Equation (4.2). The consistency of the voxelized quantities is verified in three complementary ways:

1. **Length conservation.** Summing the line-length density (see Figure 4.3a) over all pixels recovers the true curve length ℓ_{true}:

$$\sum_{I_x,I_y} \rho(I_x, I_y)\,\Delta A \approx 3.7212, \qquad \ell_{\mathrm{true}} \approx 3.7194, \qquad \text{rel. error} \approx 5.0 \times 10^{-4}.$$

   This demonstrates that voxel-based coarse graining preserves the total length.

2. **Residual check.** By construction, curvature density (Figure 4.3b) and coarse-grained curvature (Figure 4.3c) are related as

$$q(I_x, I_y) = \rho(I_x, I_y)\,\kappa_{\mathrm{cg}}(I_x, I_y),$$

   and the residual

$$r(I_x, I_y) = q(I_x, I_y) - \rho(I_x, I_y)\,\kappa_{\mathrm{cg}}(I_x, I_y)$$

   is evaluated. On a density mask ($\rho > 10^{-4}\rho_{\max}$), the residual is nearly machine zero ($L^2 \approx -4.5 \times 10^{-20}$, $L^\infty \approx 1.4 \times 10^{-17}$), confirming the self-consistency of the implementation.

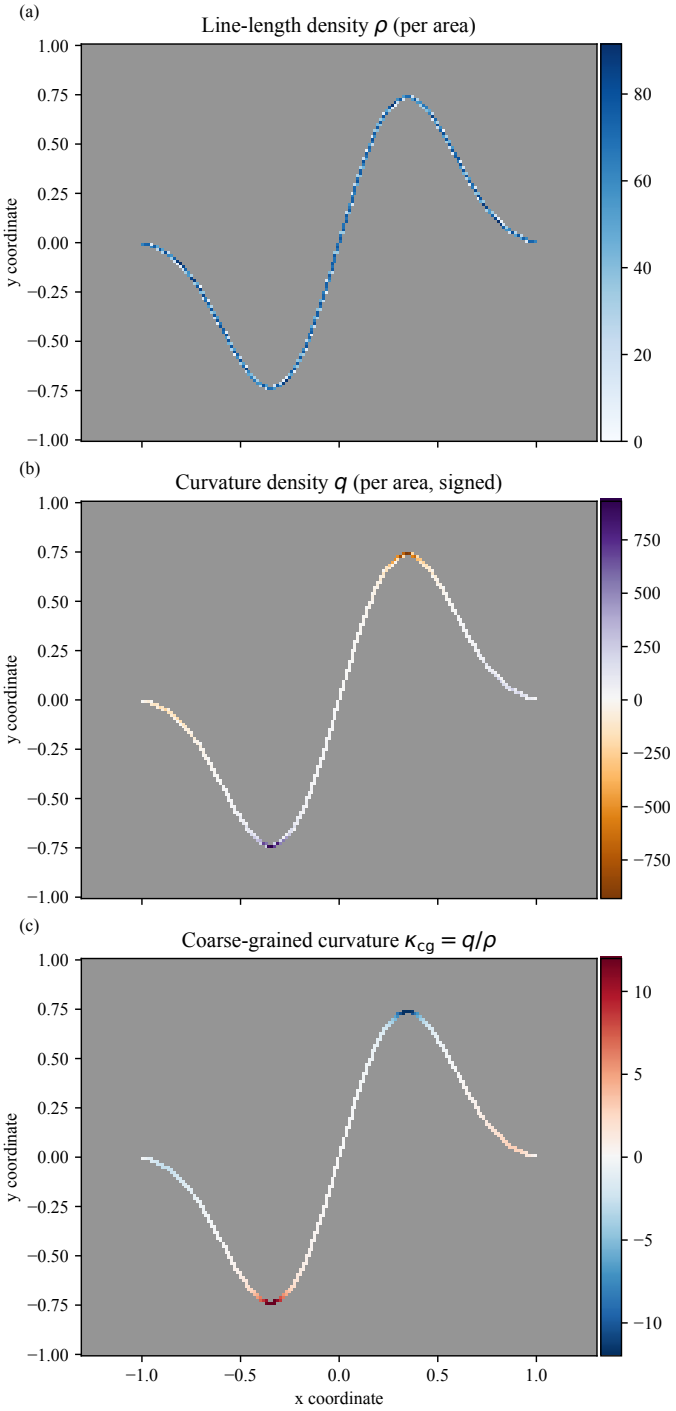

Figure 4.3. Box-filter coarse-grained fields on the S-curve (normalized by pixel area). (a) Line-length density ρ, showing length conservation $\sum \rho \Delta A \approx 3.7212$, $\ell_{\text{true}} \approx 3.7194$. (b) Curvature density q. (c) Coarse-grained curvature $\kappa_{\text{cg}} = q/\rho$, with comparison to analytic κ_{true} giving $L^2 \approx 5.12 \times 10^{-1}$, $L^\infty \approx 2.77$.

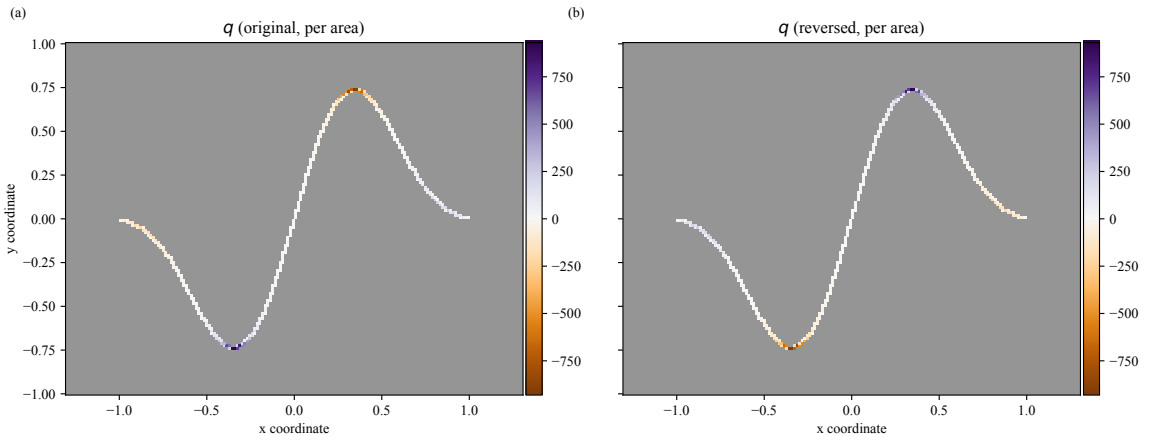

Figure 4.4. Orientation-flip validation on the S-curve. (a) Original q. (b) Flipped q, upon reversing the parametrization order, the line-length density ρ remains unchanged, while the curvature density q and coarse-grained curvature κ_{cg} correctly change sign, confirming robust handling of signed quantities.

3. **Orientation flip.** Reversing the parametrization ($u \mapsto 1-u$) leaves ρ invariant ($\sum \rho_{\mathrm{orig}}\,\Delta A \approx \sum \rho_{\mathrm{rev}}\,\Delta A \approx 3.7212$), while q and κ_{cg} change sign as expected (see Figure 4.4). On overlapping pixels, $q_{\mathrm{rev}} \approx -q_{\mathrm{orig}}$ within 5.9×10^{-4}, confirming robust treatment of signed curvature.

Finally, the grid resolution is systematically refined. Both the L^2 and L^∞ errors decrease monotonically, demonstrating the convergence of the voxel-based coarse-grained fields towards the analytic reference solution (see Figure 4.5).

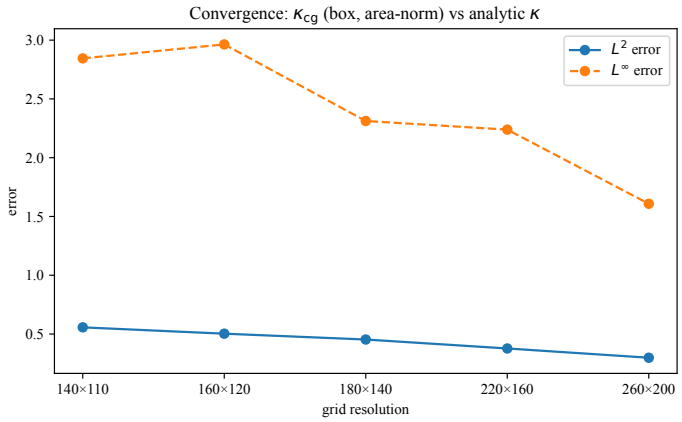

Figure 4.5. Convergence test on the S-curve. Both the L^2 and L^∞ errors decrease as the grid resolution increases.

4.3.2. Validation of Velocity Decomposition on a Semicircle

We next validate the coarse-grained velocity field on a semicircle of radius $R = 0.6$,

$$x(\theta) = R\cos\theta, \quad y(\theta) = R\sin\theta, \quad \theta \in [0, \pi],$$

with outward normal $\boldsymbol{n}(\theta) = (\cos\theta, \sin\theta)$. The coarse-grained velocity in pixel (I_x, I_y) with area ΔA is defined as

$$\boldsymbol{v}_{\mathrm{cg}}(I_x, I_y) = \frac{1}{\rho(I_x, I_y)\,\Delta A} \sum_i \boldsymbol{v}_i\, \Delta s_i\, \mathbf{1}_{\{(x_i, y_i) \in \mathrm{pixel}(I_x, I_y)\}},$$

which corresponds to the length-weighted average of all segment velocities within the pixel.

The analytic velocity corresponding to the normal law (see Figure 4.6) is tested below,

$$\boldsymbol{v}(\theta) = v\cos\theta\, \boldsymbol{n}(\theta), \qquad v = 0.4.$$

In this case, the velocity along the normal direction is $v_n(\theta) = 0.4\cos\theta$, while the tangential velocity vanishes ($v_t \equiv 0$). On a weight mask ($w > 10^{-3} w_{\max}$), the numerical errors are

$$L^2(v_t) \approx 2.4 \times 10^{-3}, \qquad L^\infty(v_t) \approx 6.8 \times 10^{-3},$$
$$L^2(v_n) \approx 2.2 \times 10^{-3}, \qquad L^\infty(v_n) \approx 5.6 \times 10^{-3}.$$

Sampling the coarse-grained velocity back onto the curve and projecting it yields the tangential and normal components (Figure 4.7a). The corresponding errors, evaluated using the L^2 and L^∞ norms defined above, remain small (Figure 4.7b), thereby confirming the accurate numerical recovery of the analytic velocity laws.

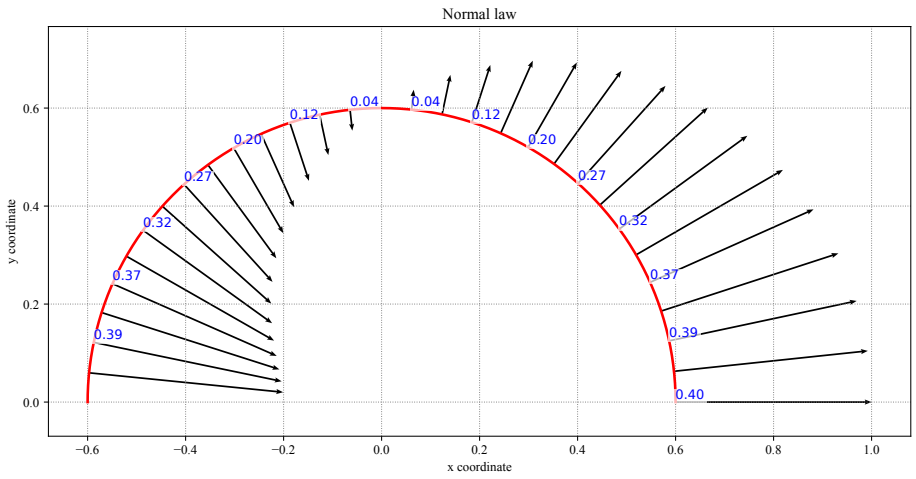

Figure 4.6. Schematic of velocity direction on a semicircle.

The velocity consistency is verified by decomposition: The case reproduces the expected $\cos\theta$ dependence of v_n with vanishing v_t. This confirms that the coarse-graining procedure respects the geometric definitions of tangential and normal velocity.

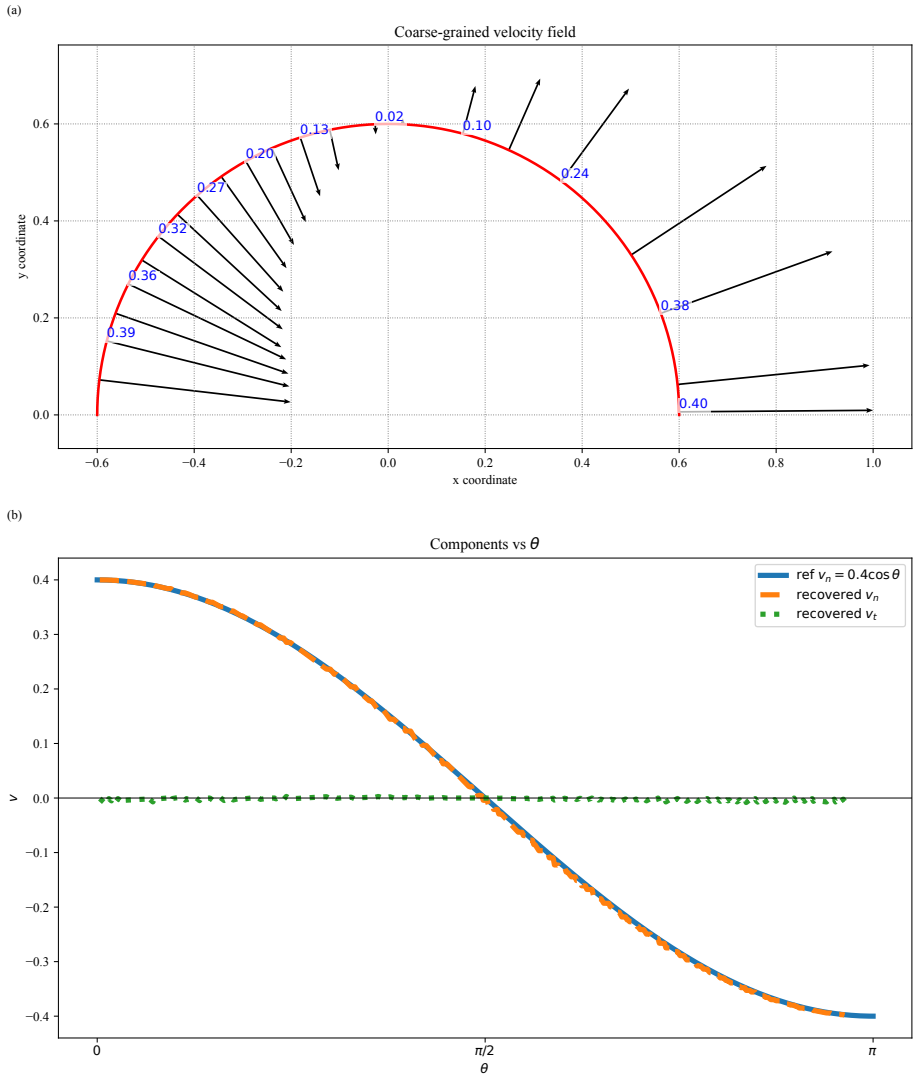

Figure 4.7. Box-filter coarse-grained velocity validation on a semicircle. (a) Coarse grained velocity field of normal law $\boldsymbol{v} = 0.4\cos\theta\,\boldsymbol{n}$. (b) The decomposition into v_t and v_n confirms correct recovery of the analytic laws.

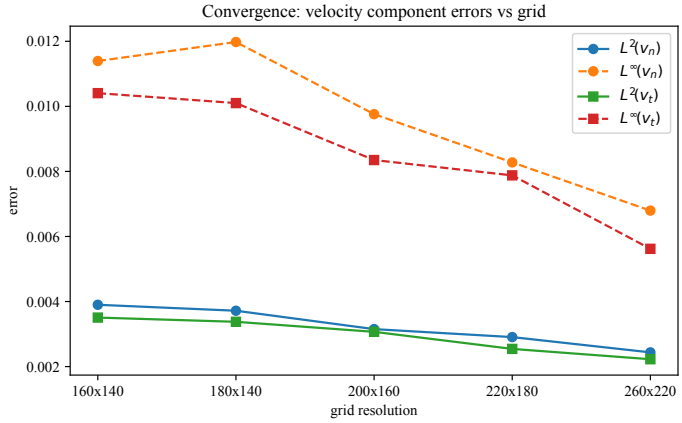

Figure 4.8. Convergence test for velocity $\boldsymbol{v} = 0.4\cos\theta\,\boldsymbol{n}$. Errors in the normal component v_n converge systematically with grid refinement. The tangential component v_t remains close to zero.

Moreover, errors decrease systematically when refining the pixel grid resolution (as shown in Figure 4.8), demonstrating convergence of the voxel-based velocity evaluation.

The test cases presented above validate the central step of the discrete-to-continuum (D2C) mapping, namely the representation of discrete dislocation segments as continuous fields of density ρ, curvature density q, signed curvature $\kappa = q/\rho$, and velocity components. These verifications establish the correctness, consistency, and convergence of the field definitions under voxel/pixel coarse graining. The codes used to perform these verifications are openly available at this repository. By contrast, the temporal coarse-graining (sliding averages over time windows) introduced in Equation (4.15) for later applications does not modify the physical definitions of the fields. It provides a complementary statistical treatment that reduces noise and improves the interpretability of experimental data. Consequently, no additional validation is required beyond the tests reported here for the spatial field representation.

4.4. Summary

In this chapter, we established a computational framework for the characterization of dislocation microstructures, motivated by the need to quantitatively interpret trajectories observed in TEM experiments. We first reviewed DDD and CDD, outlining their historical developments, methodological foundations, and intrinsic limitations. The discrete-to-continuum (D2C) approach was then introduced as a bridging methodology, providing the necessary definitions of curvature, velocity, arc length, and density tensors to translate discrete trajectories into field variables. Furthermore, the concept of spatio–temporal coarse graining (CG) was presented as a means to connect experimental observations with continuum-level descriptions. Through illustrative test cases, the consistency and applicability of the proposed framework were demonstrated.

Overall, the methods discussed here lay the groundwork for data-driven analyses of dislocation motion and provide a basis for their application in Chapter 6.

5. Probabilistic Clustering Framework for Materials Data

Unsupervised learning provides methods for discovering latent patterns in unlabeled data. In materials science, it enables the detection of hidden correlations among engineering properties and supports a probabilistic description of material classes. This capability is especially relevant for property databases such as MatWeb, where entries are high-dimensional, partially incomplete, and often exhibit class overlap.

Classical Ashby maps [165] represent an early and highly influential approach to feature engineering: two selected properties (e.g. modulus vs. density) are plotted, and classes are divided based on expert knowledge. These charts condense complex information into intuitive two-dimensional representations, but their scope is constrained by expert-defined axes, low dimensionality, and rigid categorical boundaries.

The present chapter develops a probabilistic clustering framework that extends the Ashby paradigm to contemporary materials informatics. By integrating dimensionality reduction and probabilistic clustering, the framework enables analysis in higher-dimensional spaces, automated model selection, and uncertainty-aware clustering. Rather than replacing Ashby maps, it complements them by providing a scalable, data-driven methodology for large property datasets. In Chapter 7, the same probabilistic framework is applied to nanoindentation load-displacement data of CuCr composites, where the objective shifts from separating broad material classes to inferring distinct phases within a single system.

5.1. Motivation

Unsupervised learning has become an indispensable tool for analyzing large and heterogeneous datasets. In materials science, traditional rule-based categories and manual curation are increasingly inadequate given the scale and diversity of available data [166].

Ashby maps remain an excellent example of how domain knowledge and visualization can guide material selection, yet they are constrained by dimensionality and categorical rigidity. In contrast, Gaussian mixture models offer a fully probabilistic framework: they admit anisotropic cluster shapes, provide soft assignments, and support likelihood-based model comparison through the Bayesian information criterion (BIC). These properties make GMMs particularly suited for engineering data, where material classes overlap and uncertainties are intrinsic.

The following sections outline the adopted workflow: data preprocessing, dimensionality reduction (principal component analysis (PCA)), probabilistic clustering with GMMs, model selection using BIC, and evaluation by resampling and external agreement indices, illustrated through a case study on MatWeb data.

5.2. Workflow of Unsupervised Learning

Unsupervised analysis in this work follows a four-stage pipeline (shown in Figure 5.1) tailored to engineering property data: (i) data preprocessing; (ii) dimension reduction; (iii) model selection; and (iv) model evaluation. The goal is to obtain clusters that are statistically robust and physically interpretable, while remaining consistent with the probabilistic model introduced in Section 5.3.

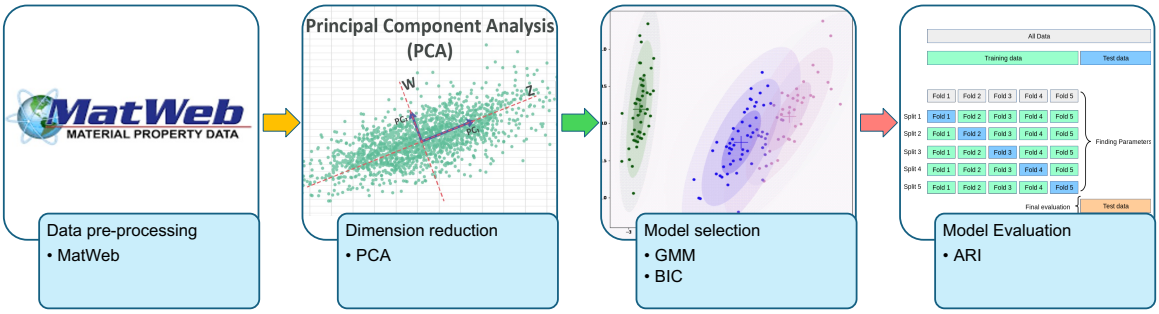

Figure 5.1. Workflow of unsupervised clustering, including preprocessing, dimensionality reduction, model selection, and evaluation.

Step 1: Data preprocessing

Raw entries from materials databases often contain inconsistent units, missing values, and range-valued reports. We therefore (a) convert all properties to SI units; (b) handle ranges by replacing them with the average of random draws within the reported bounds (and record missingness); and (c) standardize features to zero mean and unit variance. For mechanical variables used only for plotting (e.g. Ashby-type charts), log scaling is applied; modeling itself is performed on standardized variables. This prevents large-magnitude features from dominating distances and ensures numerical stability during likelihood-based covariance estimation.

Step 2: Dimensionality reduction

We use PCA to remove linear redundancy and to obtain orthogonal directions that capture the major variance. PCA is chosen here because it preserves the global variance structure required by GMMs and provides stable, deterministic low-dimensional representations.

It is important to note that PCA is a dimensionality reduction technique, not a feature selection method. While feature selection directly retains a subset of original

variables (e.g., modulus and density), PCA constructs new orthogonal variables (principal components) as linear combinations of the originals. This reduces interpretability in terms of single physical properties but improves variance preservation and numerical conditioning. In the present context, Ashby-type charts may be regarded as a form of expert-driven feature selection, whereas PCA provides a data-driven transformation that is more compatible with probabilistic clustering and BIC-based model comparison.

Step 3: Model selection

Clustering is performed with Gaussian mixture models (GMMs) (see Section 5.3). We select the model family and complexity via a grid over:

- number of components k (e.g. $k \in \{2, 3, \ldots\}$),
- covariance parameterization (full/diagonal/tied),
- retained PCA dimension q (from Step 2).

For each (k, cov, q), we compute the BIC, whose formal definition is provided in Equation (5.9), and select the model with the lowest score. This penalizes over-parameterized mixtures, which is critical when N is modest relative to d, and provides a principled way to co-tune k, covariance structure, and q.

Step 4: Model evaluation

Model evaluation depends on whether external labels are available. In the case study of this chapter, Ashby-type categories (metals, polymers, ceramics, etc.) serve as reference labels. The simplest and most transparent measure is label accuracy, i.e. the fraction of data points assigned to the correct material class. Confusion matrices are additionally used to identify systematic mislabelings (e.g. overlap between light alloys and high-performance polymers). For comparison, the adjusted Rand index (ARI) is also reported as a complementary agreement measure.

In contrast, for applications without explicit labels, such as the nanoindentation data analyzed in Chapter 7, evaluation requires indirect criteria. In that context, k-fold cross-validation is used to assess stability with respect to data splits, and the ARI quantifies agreement between different clustering assignments while discounting chance agreement.

5.3. Gaussian Mixture Model

In our context, the GMM clustering assumes that each distinct phase corresponds to a unique Gaussian distribution of Young's modulus (E) and hardness (H). The experimental dataset is denoted as

$$\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}, \quad \boldsymbol{x}_i \in \mathbb{R}^d,$$

where $\boldsymbol{x}_i$ is the ith feature vector in a d-dimensional space. The distribution of $\mathcal{D}$ can be represented by a mixture of Gaussian probability density functions (PDFs). The

resulting density is given by

$$F(\boldsymbol{x} \mid \boldsymbol{\Phi}_k) = \sum_{j=1}^{k} \alpha_j \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\theta}_j)\,, \tag{5.1}$$

where k is the number of mixture components, $\alpha_j > 0$ are the mixture weights with $\sum_{j=1}^{k} \alpha_j = 1$, $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ are the parameters of component j, and $\boldsymbol{\Phi}_k = \{\alpha_1, \ldots, \alpha_k, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$ is the complete parameter set. Here, $\boldsymbol{\mu}_j$ is the mean vector describing the location of component j, and $\boldsymbol{\Sigma}_j$ is the covariance matrix describing its spread and orientation.

5.3.1. Maximum Likelihood Estimation

Assuming that $\{\boldsymbol{x}_i\}_{i=1}^{N}$ are independent samples, the probability of observing a single data point $\boldsymbol{x}_i$ is given by

$$p(\boldsymbol{x}_i \mid \boldsymbol{\Phi}_k) = \sum_{j=1}^{k} \alpha_j \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\,. \tag{5.2}$$

The likelihood of the entire dataset $\mathcal{D}$ is then

$$\mathcal{L}(\boldsymbol{\Phi}_k) = \prod_{i=1}^{N} p(\boldsymbol{x}_i \mid \boldsymbol{\Phi}_k)\,, \tag{5.3}$$

and its log-likelihood is

$$\ln \mathcal{L}(\boldsymbol{\Phi}_k) = \sum_{i=1}^{N} \ln\left[\sum_{j=1}^{k} \alpha_j \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right]. \tag{5.4}$$

Since $\mathcal{L}$ can take extremely small values for large N, the logarithm is used for numerical stability. Maximizing $\ln \mathcal{L}$ with respect to $\boldsymbol{\Phi}_k$ yields the optimal parameter estimates.

5.3.2. Expectation-Maximization Algorithm

Direct maximization of $\ln \mathcal{L}$ leads to a system of coupled nonlinear equations. The standard approach is the expectation-maximization (EM) algorithm. Define the posterior responsibility of component j for data point $\boldsymbol{x}_i$ as

$$\gamma_{ij} = p(j \mid \boldsymbol{x}_i) = \frac{\alpha_j \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}{\sum_{l=1}^{k} \alpha_l \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{\mu}_l, _l)}. \tag{5.5}$$

Then the EM updates are then given by

$$\boldsymbol{\mu}_j^{\text{new}} = \frac{\sum_{i=1}^{N} \gamma_{ij}\, \boldsymbol{x}_i}{\sum_{i=1}^{N} \gamma_{ij}}, \tag{5.6}$$

$$\boldsymbol{\Sigma}_j^{\text{new}} = \frac{\sum_{i=1}^{N} \gamma_{ij} \, (\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top}{\sum_{i=1}^{N} \gamma_{ij}}, \tag{5.7}$$

$$\alpha_j^{\text{new}} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij}, \tag{5.8}$$

where $(\cdot)^\top$ denotes matrix transpose. Note that in the one-dimensional special case, the outer product $(\boldsymbol{x}_i - \boldsymbol{\mu}_j)(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^\top$ reduces to the scalar square $(x_i - \mu_j)^2$. The E-step computes γ_{ij}, and the M-step updates the parameters according to Equations (5.6) to (5.8). Iterating E and M steps monotonically increases the log-likelihood until convergence to a local maximum.

5.3.3. Bayesian Information Criterion

To compare models with different numbers of components k, we employ the BIC. It is well established and commonly applied in model selection [167], providing a measure for assessing the accuracy of the unsupervised GMM. For maximum-likelihood-based models, BIC is defined as

$$\text{BIC} = -2 \ln \hat{\mathcal{L}} + d \ln N. \tag{5.9}$$

Here, d is the number of free parameters, N is the sample size, and $\hat{\mathcal{L}}$ is the maximized likelihood. The preferred model is the one with the smallest BIC. The penalty term $d \ln N$ discourages overfitting by favoring simpler models unless the increase in likelihood justifies additional complexity.

5.4. Model Evaluation

Model validation is crucial to ensure the stability, generalizability, and reliability of clustering results. Unlike supervised learning, where performance can be directly assessed against known labels, unsupervised evaluation requires strategies that focus on the consistency and robustness of the derived structures. Beyond algorithmic performance, it is necessary to assess clustering stability, detect potential under- or overfitting, and compare competing models. In this work, we employ k-fold cross-validation in combination with external agreement measures to evaluate the quality of clustering outcomes.

5.4.1. k-fold Cross-Validation

Although k-fold cross-validation is conventionally applied in supervised learning, here we adopt it as a resampling strategy to assess the stability of clustering results and the sufficiency of the available data. In this context, it is not used as a predictive validation tool but rather as a means to examine the robustness of the GMM assignments across different data splits.

Let $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^N$ be the dataset, with associated cluster labels $\{y_i\}_{i=1}^N$. A clustering algorithm A_k produces an assignment $\hat{y}_i := A_k(\boldsymbol{x}_i)$. In our case, the features are Young's modulus (E) and hardness (H), and $\hat{y}_i$ are obtained from the optimal GMM with three clusters.

In k-fold cross-validation, $\mathcal{D}$ is split into k equal parts. Each fold is used once as the validation set, while the remaining $k-1$ folds form the training set. The process is repeated k times, and performance is averaged across folds. The Rand index is used to assess the consistency of assignments across folds.

5.4.2. Adjusted Rand Index

In the absence of ground-truth labels, the evaluation of unsupervised models relies on measures that quantify the consistency of cluster assignments rather than predictive accuracy. The Adjusted Rand Index (ARI) provides such a measure by comparing the similarity between two partitions of the same dataset while accounting for random agreement. Here, ARI is used to quantify the stability of GMM clustering results across different cross-validation folds.

Let $\mathcal{C} = \{C_1, \ldots, C_r\}$ denote the ground-truth partition and $\mathcal{K} = \{K_1, \ldots, K_s\}$ the clustering result. For all $\binom{N}{2}$ possible pairs of elements in the dataset:

- a: number of pairs that are placed in the same subset in both $\mathcal{C}$ and $\mathcal{K}$,
- b: number of pairs that are placed in different subsets in both $\mathcal{C}$ and $\mathcal{K}$.

The unadjusted Rand index between $\mathcal{C}$ and $\mathcal{K}$ is given by

$$\mathrm{RI}(\mathcal{C}, \mathcal{K}) = \frac{a+b}{\binom{N}{2}}. \tag{5.10}$$

Here, $\binom{N}{2}$ is the total number of pairs.

However, RI does not guarantee that random cluster assignments yield values close to zero, especially when the number of clusters is large. To correct for this, the ARI is defined as

$$\mathrm{ARI}(\mathcal{C}, \mathcal{K}) = \frac{\mathrm{RI}(\mathcal{C}, \mathcal{K}) - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}. \tag{5.11}$$

Here, $\mathbb{E}[\mathrm{RI}]$ is the expected Rand index under random labelings. ARI ranges from 0 (random agreement) to 1 (perfect agreement).

5.5. Case Study: Materials Clustering Using Gaussian Mixture Models

In this case study, we apply the probabilistic clustering framework introduced in this chapter to engineering property data from MatWeb, using Ashby's labels as a benchmark. The objective is not to reproduce Ashby charts exactly, but to test whether Gaussian mixture modeling can recover comparable groupings automatically, and to identify regions of overlap or ambiguity.

5.5.1. Data Collection and Preprocessing

Ashby's framework distinguishes nine material classes using ten properties [165]. In practice, the MatWeb database [168] provides overlapping but incomplete coverage, with 437 entries spanning five categories and seven properties, as indicated in Table 5.1. While this accounts for over 90% of the Ashby categories, MatWeb uses a distinct nomenclature that merges or subdivides some classes. Missing values are frequent for certain properties, reflecting the heterogeneous nature of materials databases.

Table 5.1. Overview of Collected Data by Material Class

		Feature						
Class	No.	D	E	TC	CTE	TS	Eab	ER
Metal	60	60	60	60	60	60	58	50
Polymer	216	216	118	151	162	178	171	148
Ceramic	131	131	117	128	117	52	1	49
Wood & Natural Material	21	21	21	21	-	16	-	-
Other Engineering Material	9	9	9	9	-	9	-	-
In total:	437	437	325	369	339	315	230	247

Notes: No. = Number of observations in each class; D = Density; E = Modulus of Elasticity; TC = Thermal Conductivity; CTE = Coefficient of Thermal Expansion; TS = Tensile Strength; Eab = Elongation at Break; ER = Electrical Resistivity.

Preprocessing followed the same procedure described in Section 5.2: units were converted to SI, range values were averaged over random draws within their bounds, and features were standardized. All processed data and analysis scripts are openly available at this repository.

As shown in Figure 5.2, the overall trends of the MatWeb-based scatter plot reproduce the groupings familiar from Ashby's modulus-density chart: polymers occupy the low-density, low-modulus region; metals concentrate at higher densities and moduli; and ceramics form the upper-right envelope. The key difference lies in representation: Ashby maps smooth local variability into class envelopes, whereas the database plot directly reflects heterogeneous entries and outliers. By fitting covariance-based ellipses in log-density–log-modulus space for each class and overlaying their projections on the scatter plot, a data-driven analogue of Ashby's envelopes is obtained. The tilt and extent follow the principal directions of variability, while preserving the raw distribution and its outliers. This comparison highlights the motivation for probabilistic clustering: to provide automated group detection and uncertainty quantification in contexts where boundaries are not predefined.

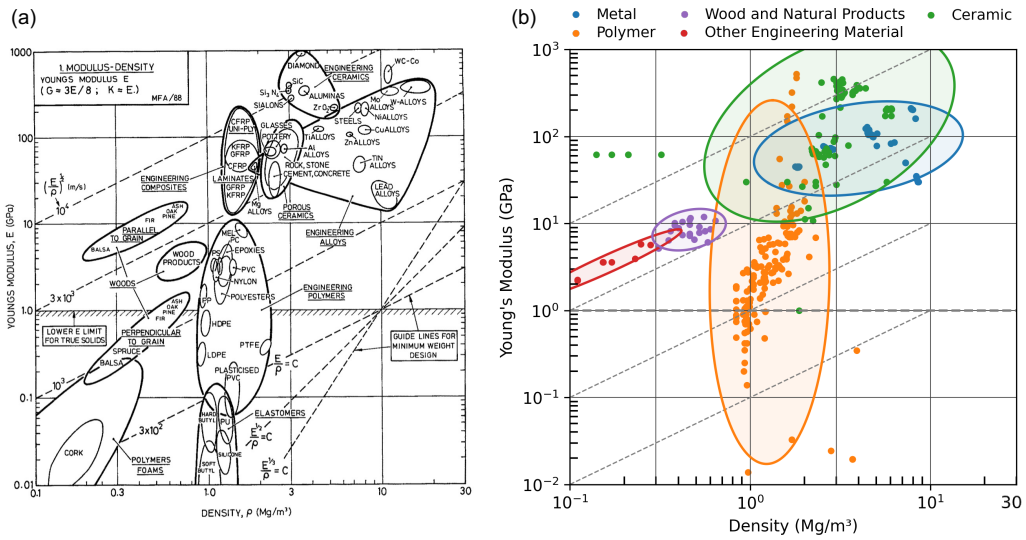

Figure 5.2. Comparison of Ashby's modulus–density chart and a MatWeb-based plot. (a) Ashby's chart (adapted from [165]). (b) MatWeb-based plot. While the overall groupings are consistent, the database view reveals heterogeneity and overlaps that motivate probabilistic clustering.

5.5.2. Principal Component Analysis

Among the 437 MatWeb entries, only 136 contained complete information across all seven features (density (D), modulus of elasticity (E), tensile strength (TS), thermal conductivity (TC), coefficient of linear expansion (CTE), electrical resistivity (ER), and elongation at break (Eab)), and these records belong exclusively to two categories: Metals and Polymers. This highlights the sparsity of complete data as dimensionality increases. To address the issue of limited samples in a high-dimensional space, we apply principal component analysis (PCA) to reduce the feature space while preserving the dominant variance structure.

The eigenvalue spectrum of the PCA with two-class is shown in Figure 5.3a. PC1 explains 49.2% of the total variance, PC2 adds 19.0%, and PC3 a further 12.1%, so that the first three components already capture 80.3% of the dataset variability. Including PC4 and PC5 increases the cumulative variance explained to 96.5%, while the remaining two components contribute less than 3%. These results indicate that a reduced feature space of three to five principal components suffices to retain the essential variation, facilitating clustering and visualization with minimal information loss. For the three-class analysis, the same procedure was applied. PCA was performed on the six available features (D, E, TS, TC, CTE, and ER), and the resulting loading patterns were used to guide the construction of feature subsets (see Figure 5.3b). These subsets form the basis for the clustering models developed in Section 5.5.4.

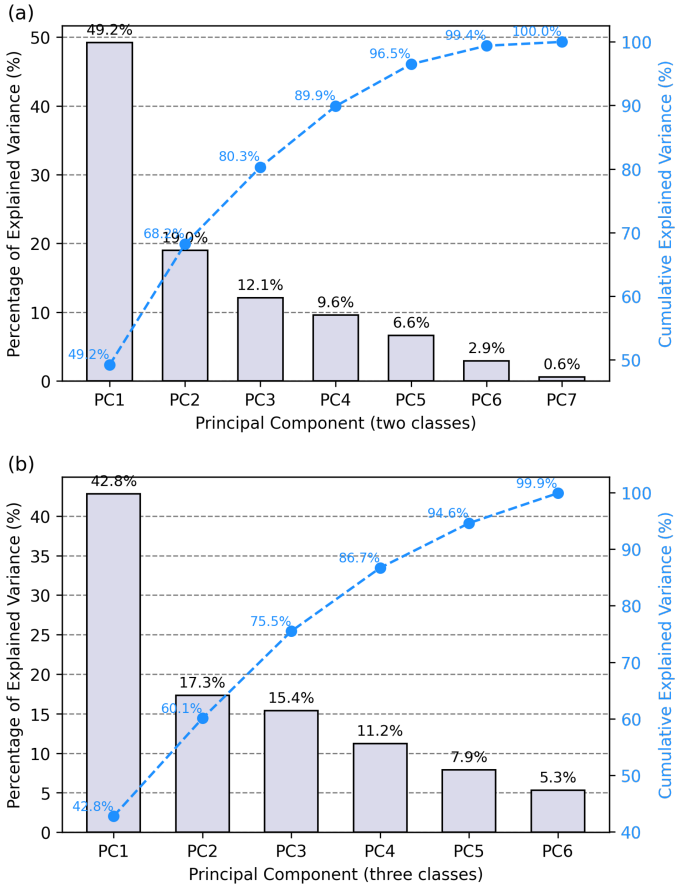

Figure 5.3. Scree plot with cumulative explained variance for different class and dimensionality settings. (a) Two classes in seven dimensions. (b) Three classes in six dimensions.

5.5.3. Two Classes: From Seven to One Dimension

The visualization of the GMM results, using most features as a benchmark, is shown in Figure 5.4 and achieves an accuracy of 97%. The PCA projections demonstrate that the two categories, Metal and Polymer, are largely separable in the reduced two-dimensional space. Nevertheless, the confusion matrix reveals that, while all Metal samples (49/49, 100%) were correctly classified, a small subset of Polymer samples (4/87, 4.6%) were misclassified as Metal. A manual inspection of these four mislabelled polymers showed that the errors stem from atypical property combinations: in some cases, the (E, TC) pair far exceeded the typical polymer range (with TC values approaching those of metals), while in others, the (E, TS) pair was several times higher than the polymer average. By contrast, polymers that exhibited an extreme value in only a single property while remaining otherwise similar to the broader polymer class were still correctly recognized as polymers, indicating that mislabelling arises mainly from multi-property deviations rather than isolated anomalies. These results highlight both the effectiveness of the GMM in capturing the major class structure and the inherent challenges of achieving perfect separation in complex material datasets.

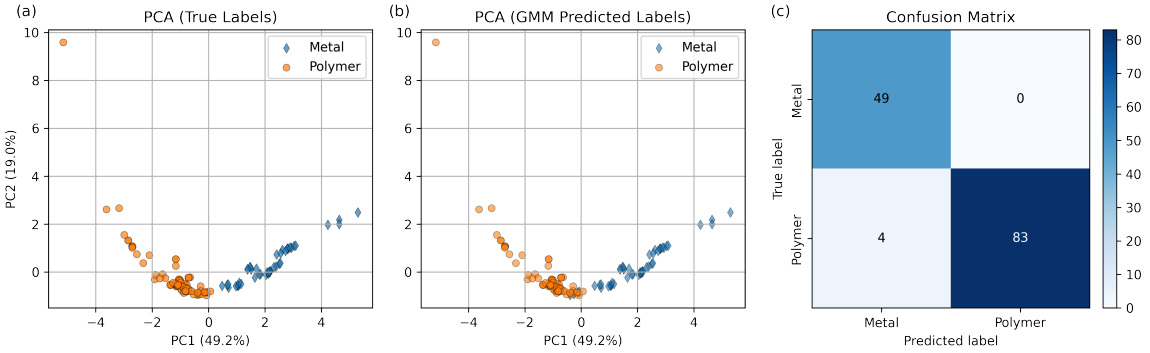

Figure 5.4. Results of the GMM clustering. (a) PCA projection with true labels (Metal vs. Polymer). (b) PCA projection with GMM-predicted clusters. (c) Confusion matrix showing high overall accuracy 97.06%, with perfect clustering for Metal and minor misclustering in the Polymer group.

Based on the PCA coefficient matrices of two classes(Figure 5.3a), the first three principal components - which together accounted for over 80.3% of the total variance, were dominated by E, TS, and D, with secondary contributions from CTE and ER. In contrast, TC contributed relatively little to the variance structure, while Eab was only significant for the second component. These results guided the feature selection strategy: focusing on E, TS, and D as the core subset, with optional inclusion of CTE or ER to capture additional thermal or electrical information.

As summarized in Table 5.2, six models were designed to systematically evaluate the effect of feature selection on clustering performance. The baseline model (M1) uses all seven available features (D, E, TS, TC, CTE, ER, and Eab), achieving high accuracy but with a limited number of samples (136). To examine the minimal requirements for effective clustering, models with reduced feature sets were constructed. M6 includes

only a single mechanical property (E), while M5 adds TS to form a core two-feature representation of mechanical strength. M4 further incorporates D, representing the three dominant features highlighted by the PCA coefficient matrices. This subset is expected to approach the performance of the full feature set while significantly increasing data availability. To evaluate the role of cross-domain physical properties, two four-feature variants were introduced: M3a (E, TS, D, ER) emphasizes electrical resistivity, whereas M3b (E, TS, D, CTE) emphasizes thermal expansion. These two models allow direct comparison of the discriminative power of electrical versus thermal properties when combined with core mechanical descriptors. Finally, M2 combines both ER and CTE with the three core mechanical features, providing a balanced model that retains essential mechanical, structural, and physical information, while still enabling a larger dataset than the full seven-feature model. The corresponding clustering results for Model M2 - Model M5 are illustrated in Figure 5.5, Figure 5.6, Figure 5.7, Figure 5.8, and Figure 5.9, respectively.

Table 5.2. Summary of two-class GMM model design and evaluation results

Model	Selected Features	No.	Wrong	Accuracy	BIC	ARI
M1	D, E, TS, TC, CTE, ER, Eab	136	4	97.06%	-284.68	0.88
M2	E, TS, D, ER, CTE	153	12	92.16%	30.49	0.71
M3a	E, TS, D, ER	170	17	90.00%	428.68	0.64
M3b	E, TS, D, CTE	193	12	93.78%	627.01	0.76
M4	E, TS, D	226	20	91.15%	-1.06	0.67
M5	E, TS	226	17	92.48%	57.70	0.71
M6	E	248	20	91.94%	17.12	0.69

Notes: BIC is reported for completeness, although the number of clusters was fixed a priori and BIC was not used for model selection in this study.

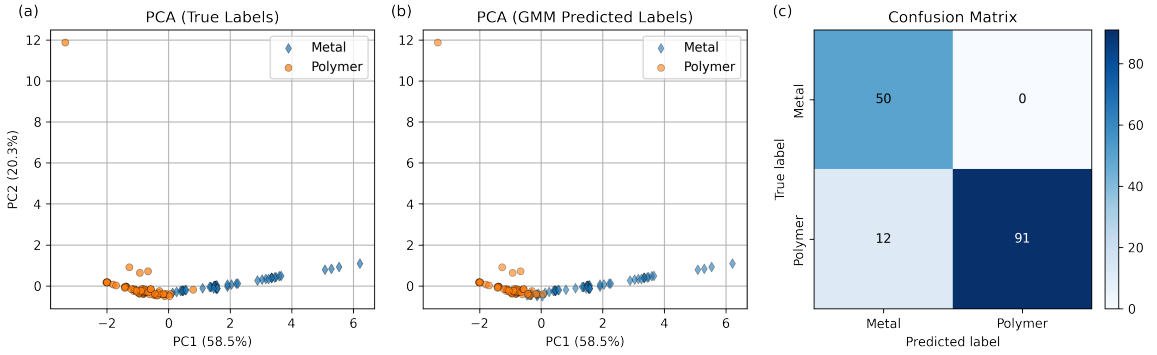

Figure 5.5. Clustering results obtained using PCA and GMM with model M2. The analysis involves two categories in a five-dimensional feature space, based on 153 samples, and achieves an accuracy of 92.16%.

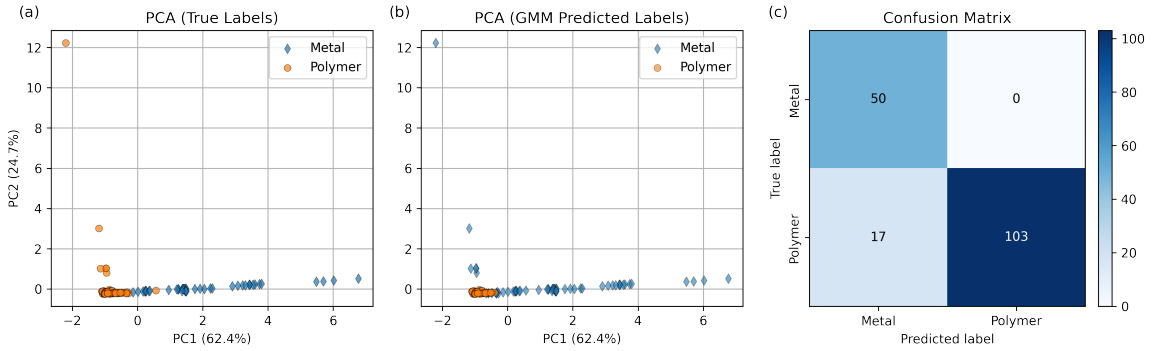

Figure 5.6. Clustering results obtained using PCA and GMM with model M3a. The analysis involves two categories in a four-dimensional feature space, based on 170 samples, and achieves an accuracy of 90.00%.

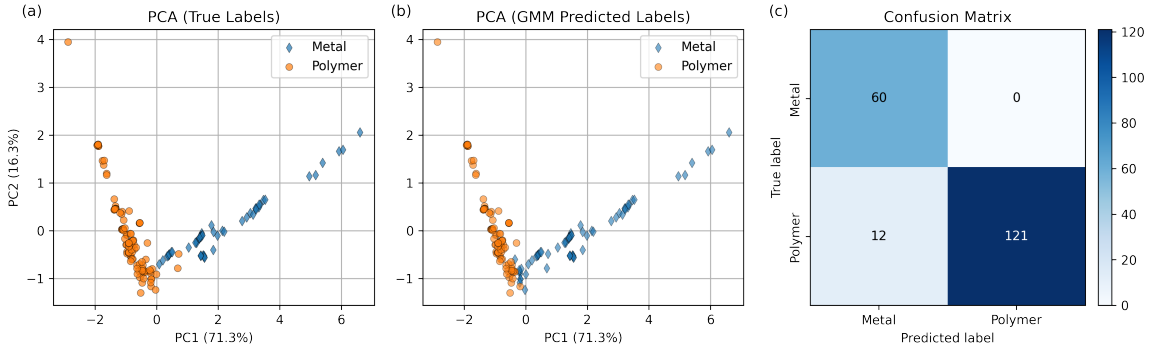

Figure 5.7. Clustering results obtained using PCA and GMM with model M3b. The analysis involves two categories in a four-dimensional feature space, based on 193 samples, and achieves an accuracy of 93.78%.

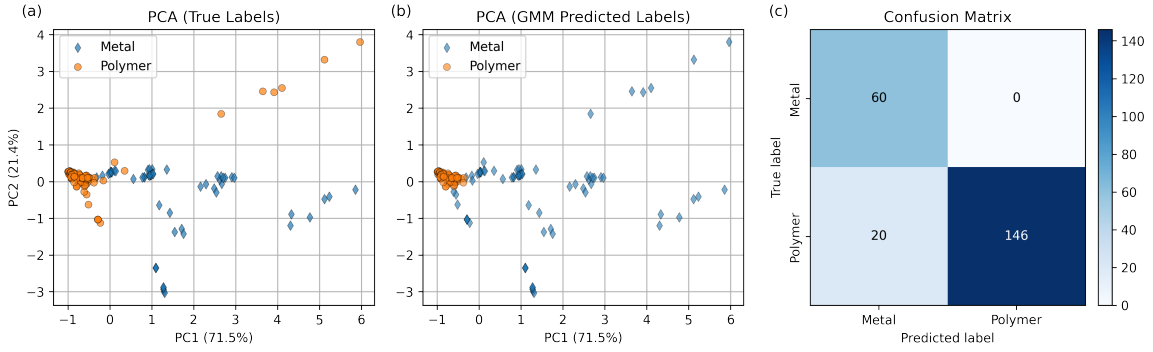

Figure 5.8. Clustering results obtained using PCA and GMM with model M4. The analysis involves two categories in a three-dimensional feature space, based on 226 samples, and achieves an accuracy of 91.15%.

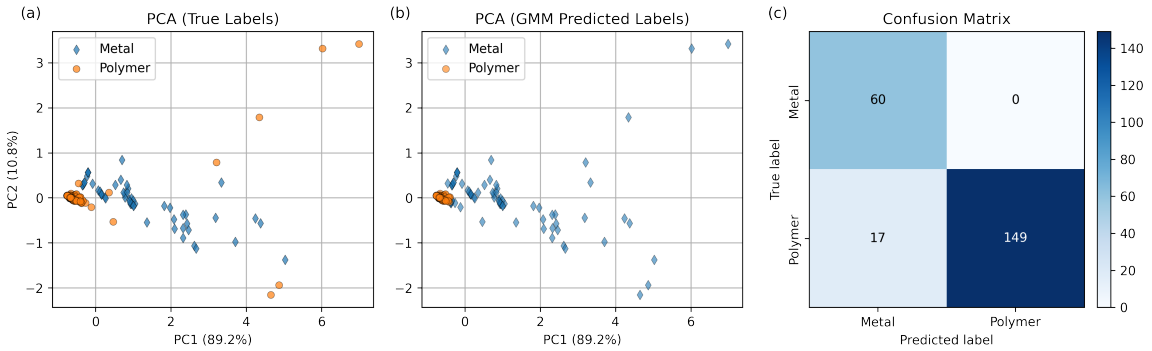

Figure 5.9. Clustering results obtained using PCA and GMM with model M5. The analysis involves two categories in a two-dimensional feature space, based on 226 samples, and achieves an accuracy of 92.48%.

The evaluation results in Table 5.2 demonstrate that high clustering accuracy can be achieved even with a reduced number of features. The baseline model (M1), using all seven features, achieved the highest accuracy (97.06%) but was limited to 136 samples. Remarkably, models with fewer features maintained comparable performance: for instance, M6 with only a single feature (E) reached 91.94% accuracy, while M5 (E and TS) and M4 (E, TS, D) both exceeded 91%. This confirms that a small subset of mechanically meaningful features is sufficient to separate metals and polymers.

The comparison between M3a and M3b further highlights the role of feature type. Although both models used four features, the inclusion of electrical resistivity (M3a) resulted in relatively lower performance (90.00%, ARI = 0.64), whereas the inclusion of thermal expansion (M3b) preserved high accuracy (93.78%, ARI = 0.76). Finally, the balanced model M2, combining both electrical and thermal properties with the three core features, achieved stable performance (92.16%) with a moderate sample size.

Overall, these results confirm that the representativeness of selected features plays a more critical role than the sheer number of features, and that carefully chosen subsets can retain high accuracy while substantially increasing data availability.

5.5.4. Three Classes: From Six to One Dimension

Figure 5.10 illustrates the performance of the three-class GMM model trained on six features. The PCA projection in Figure 5.10a reveals that polymers occupy a compact and well-separated region, whereas metals and ceramics are positioned closer together with visible overlap. The corresponding GMM clustering in Figure 5.10b largely reproduces this trend, assigning polymers to a distinct cluster while mixing ceramics and metals in the overlapping region. Figure 5.10c is the confusion matrix that quantifies these observations: metals are classified with perfect accuracy (50/50, 100%), polymers are recognized with high reliability (87/93, 93.5%), but ceramics are frequently misclassified as metals (14/26, 53.8%). The overall accuracy reaches

89.35%, indicating that the model effectively separates polymers and reliably identifies metals, but still struggles to resolve the boundary between metals and ceramics.

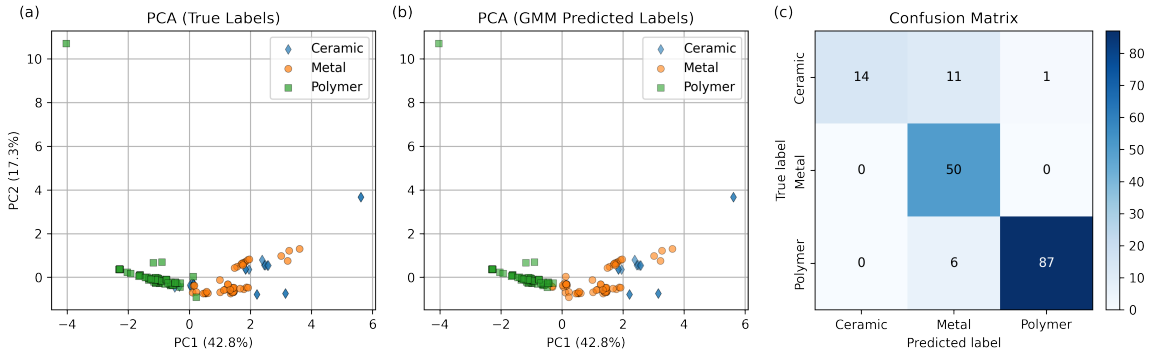

Figure 5.10. Results of the GMM clustering. (a) PCA projection with true labels (Metal vs. Polymer vs. Ceramic). (b) PCA projection with GMM-predicted clusters. (c) Confusion matrix showing high overall accuracy. For illustration, a PCA visualization in 3D is provided in Figure C.1; all subsequent PCA plots are shown in 2D to ensure clarity of labeling.

Given the PCA coefficient matrices of three classes (Figure 5.3b), since E, TS, and D dominated the first principal component, they were chosen as the core features. Additional models were designed by incrementally adding either ER or CTE to test the role of electrical and thermal properties, respectively, followed by a balanced model that combines both. The baseline model includes all six features. This hierarchical design enables a systematic evaluation of how feature representativeness and number of features jointly affect clustering performance under the three-class scenario.

Table 5.3 summarizes the clustering performance for models with three target classes (metals, polymers, ceramics). The results show that feature-rich models (e.g., m1 with six attributes) achieve relatively high accuracy (89.35%) and a strong adjusted Rand index (ARI = 0.75), indicating that the chosen features contribute effectively to class separation. Reducing the feature set generally lowers performance, although some intermediate models such as m3a (four features) still retain a competitive accuracy close to 90%. By contrast, minimal feature sets (m5: two features, m6: one feature) result in a clear decline in clustering quality, with accuracy falling to around 76%–80% and ARI values around 0.56 - 0.61. This trend highlights both the importance of multi-property information and the sensitivity of the Gaussian mixture model to feature dimensionality in the three-class setting. The corresponding clustering results for Model m2 - Model m4 are illustrated in Figure 5.11, Figure 5.12, Figure 5.13, and Figure 5.14, respectively.

When comparing Table 5.3 and Table 5.2, a clear distinction emerges between the three-class and two-class clustering tasks. For the binary case (metals vs. polymers), accuracies consistently exceed 90% across a wide range of feature subsets, with the best model (M1) reaching 97% accuracy and ARI = 0.88. In contrast, the three-class case shows lower overall performance, with accuracies mostly around 80-89% and

Table 5.3. Summary of three-class GMM model design and evaluation results

Model	Selected Features	No.	Wrong	Accuracy	BIC	ARI
m1	D, E, TS, TC, CTE, ER	169	18	89.35%	-482.40	0.75
m2	E, TS, D, ER, CTE	179	35	80.45%	-201.50	0.65
m3a	E, TS, D, ER	242	25	89.67%	460.74	0.72
m3b	E, TS, D, CTE	196	38	80.61%	-621.16	0.50
m4	E, TS, D	276	51	81.52%	475.35	0.62
m5	E, TS	276	56	79.71%	-141.00	0.61
m6	E	365	86	76.43%	130.39	0.56

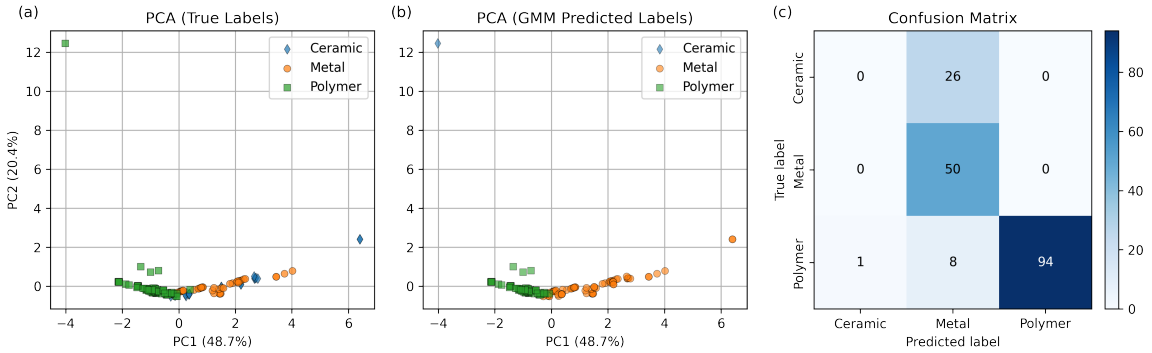

Figure 5.11. Clustering results obtained using PCA and GMM with model m2. The analysis involves three categories in a five-dimensional feature space, based on 179 samples, and achieves an accuracy of 80.45%.

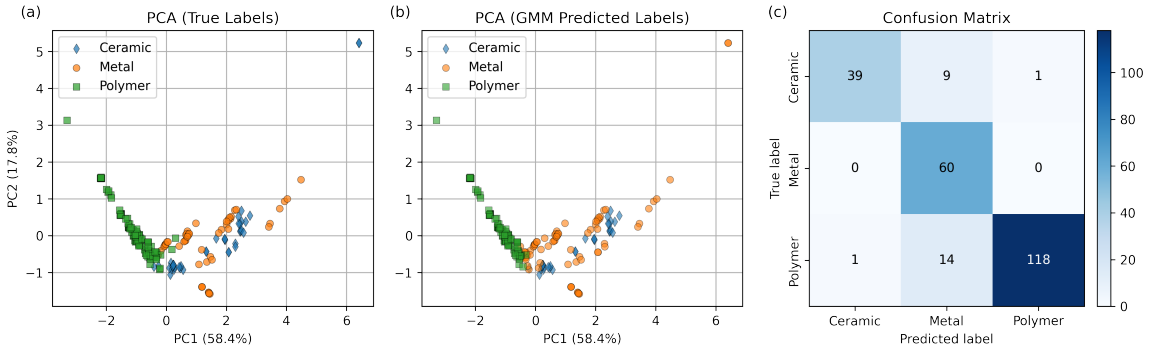

Figure 5.12. Clustering results obtained using PCA and GMM with model m3a. The analysis involves three categories in a four-dimensional feature space, based on 242 samples, and achieves an accuracy of 89.67%.

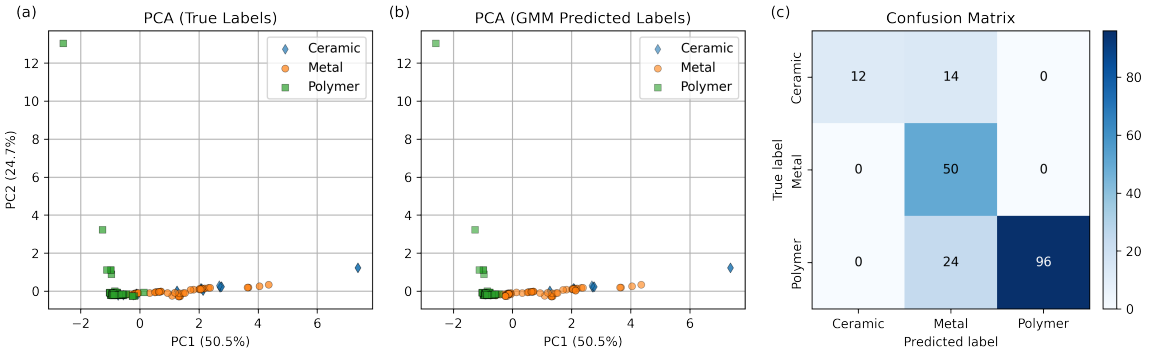

Figure 5.13. Clustering results obtained using PCA and GMM with model m3b. The analysis involves three categories in a four-dimensional feature space, based on 196 samples, and achieves an accuracy of 80.61%.

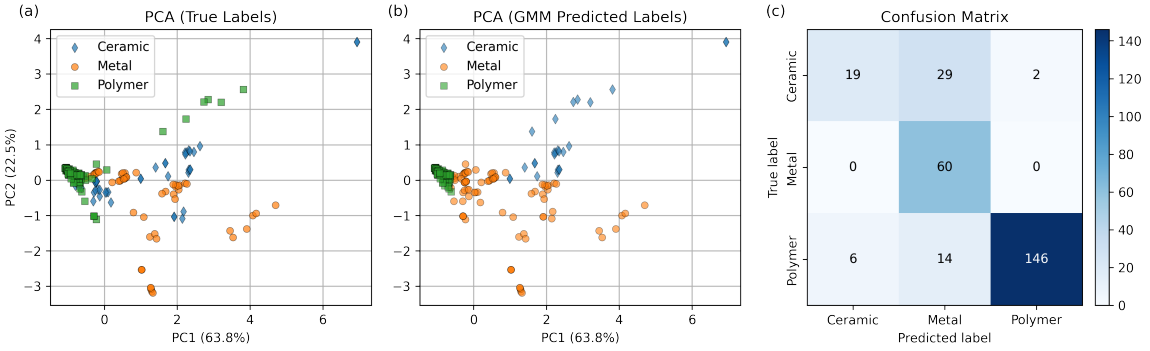

Figure 5.14. Clustering results obtained using PCA and GMM with model m4. The analysis involves three categories in a three-dimensional feature space, based on 276 samples, and achieves an accuracy of 81.52%.

ARI values rarely exceeding 0.75. This difference reflects the intrinsic difficulty of separating ceramics from metals and polymers, as their property ranges overlap more substantially. The comparison underscores a general point: binary separations of broad material classes can be robustly captured even with reduced feature sets, while multi-class clustering requires richer descriptors and remains more prone to ambiguity.

It is also worth noting that accuracy and ARI do not always exhibit perfect agreement in Table 5.3. For example, models m1 and m3a achieve similar accuracies close to 90%, yet their ARI values differ slightly, while m3b shows a much lower ARI despite a still moderate accuracy. This discrepancy likely reflects the fact that the MatWeb class labels are not strict ground truth but database-assigned categories that can contain overlaps or ambiguities, particularly between metals and ceramics. In such cases, the ARI provides a more robust measure of clustering quality, as it evaluates the overall consistency of pairwise assignments independently of arbitrary label permutations. Hence, ARI may better capture the intrinsic structure of the data, whereas accuracy remains tied to the limitations of the database labels.

5.5.5. Special Case: Two Classes with Four Dimensions

As a special case, clustering was also tested on a small subset of only 25 samples comprising two categories: other engineering materials and wood-based natural products (Figure 5.15). Here the classes are almost perfectly separated in the PCA projection, and the GMM achieved 100% accuracy. Although this result lacks statistical significance due to the very limited sample size, it illustrates that when class differences are inherently large, even simple models can provide flawless separation.

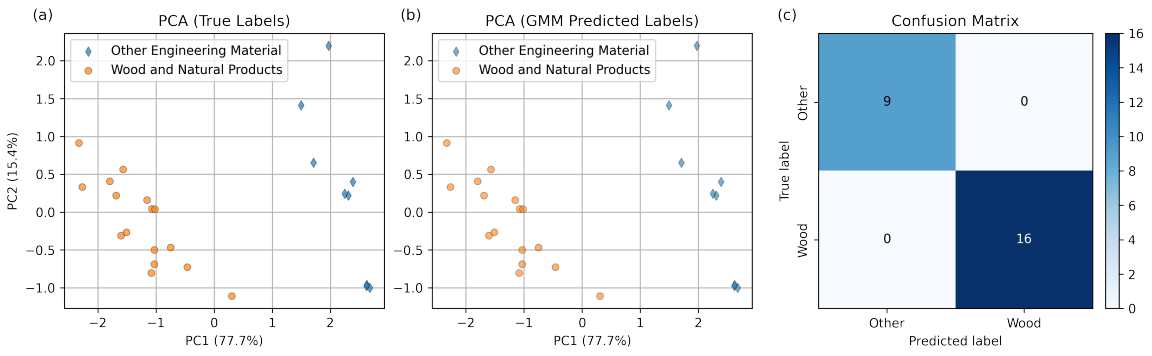

Figure 5.15. Comparison between the data distribution and clustering results. (a) Distribution of filtered data. (b) Optimal clustering results distinguishing wood from other engineering materials, achieving 100% accuracy.

Discussion Overall, the case study shows that probabilistic clustering can recover the main groupings familiar from Ashby charts, while also exposing regions of overlap between material classes. It is important to emphasize that clustering is not equivalent to supervised classification: the algorithm itself does not learn from labels, but rather discovers group structure directly from the data. The reference categories from Ashby

maps serve here only as an external benchmark, allowing us to evaluate how closely the data-driven clusters align with expert-defined classes.

This distinction highlights both the strength and the limitation of unsupervised approaches. On the one hand, clustering provides a way to generalize Ashby's two-dimensional visualizations into higher-dimensional, probabilistic groupings. On the other hand, discrepancies between clusters and reference categories remind us that material boundaries are not always sharp, and that overlaps are intrinsic to the heterogeneous nature of property datasets. These insights set the stage for Chapter 7, where the same framework is applied to a different context: nanoindentation data without predefined labels. The goal is to resolve phases within a single material system.

5.6. Summary

This chapter established a probabilistic clustering framework for materials data. By extending the principle of Ashby maps into high-dimensional property spaces, GMMs enable automatic, probabilistic grouping and explicit quantification of uncertainty. Case studies on MatWeb data demonstrate both the promise (accurate recovery of canonical separations) and the limitations (reduced accuracy in overlapping regions due to database heterogeneity).

Importantly, the methodological framework developed here is not confined to property charts: in the following Chapter 7, we apply it to nanoindentation load-displacement data, where probabilistic clustering reveals latent deformation modes and supports data-driven predictions of mechanical performance.

Part III.

APPLICATIONS AND RESULTS

6. Data-mining of *in situ* TEM Experiments: on the Dynamics of Dislocations in CoCrFeMnNi Alloys

6.1. Introduction

This chapter demonstrates the applicability of the 3D reconstruction and D2C framework developed in Chapter 3 and Chapter 4 in real *in situ* TEM experiments. To showcase its practical use, we conducted an examination of a representative pile-up of dislocations that were moving along closely packed planes in a Cantor (CoCrFeMnNi equiatomic) alloy. Subsequently, we reconstructed the three-dimensional microstructures of the dislocations, enabling us to directly see the movement of dislocations on a specific glide plane. Next, a simple model was developed to quantitatively assess the force exerted by the pinning point and its subsequent changes, utilizing the bent angle of dislocation as a basis. In the last section, we carried out a spatio-temporal analysis of the localized pinning points by employing a coarse-graining approach to the dislocation microstructure, akin to the traditional scanning technique [169]. The present study demonstrates that dislocation lines encounter a dynamic landscape as the deformation proceeds and that these local obstacles are probably at the root of hardening mechanisms in the Cantor alloy. The method proposed in this study has the potential to be extended to various MCCA in order to evaluate the average "friction stress" that would arise from these interactions. The content of this chapter has been published in [164].

6.2. Methods

Below, we explain all the methods utilized in the analysis in detail. Section 6.2.1 describes how the experiments were performed. Section 6.2.2 introduces how the pinning point strength can be estimated, and Section 6.3.3 applies the coarse-graining approach introduced in Section 4.2.3 to analyze the spatio-temporal evolution of dislocation microstructures. The three-dimensional reconstruction of dislocation lines from two-dimensional TEM images, which forms the second step after the experiment, was already presented in Section 3.3.2 and is only briefly referred to here.

6.2.1. *In situ* TEM Straining Experiment

In situ TEM straining experiments were conducted on a Cantor alloy prepared at Mines Saint-Etienne, France. Electron-transparent, stretchable specimens were fabricated following the procedure described in the previous work [130]. Additional videos illustrating the *in situ* TEM experiments are available online. The videos were typically recorded after incremental strain steps were applied to the sample using a Gatan model 671 straining holder operated at 96 K. They were captured with a Megaview III SIS CCD camera and directly stored in MPEG4 format on a dedicated hard drive.

6.2.2. Estimation of Pinning Point Strength

In this section, we utilize the simple line tension model to estimate the strength of the pinning point. The line tension of a dislocation segment with mixed character is given by

$$T = E_{\mathrm{el}}(\alpha) + \frac{\mathrm{d}^2 E_{\mathrm{el}}(\alpha)}{\mathrm{d}\alpha^2}. \tag{6.1}$$

Here, α is the angle between the Burgers vector and the line direction, and $E_{\mathrm{el}}(\alpha)$ is the elastic energy of the dislocation, which can be written as [12]

$$\begin{aligned} E_{\mathrm{el}}(\alpha) &= \left[\frac{Gb^2 \sin^2\alpha}{4\pi(1-\nu)} + \frac{Gb^2\cos^2\alpha}{4\pi}\right] \ln\left(\frac{R_{\mathrm{ext}}}{r_0}\right) \\ &= \frac{Gb^2\left(1-\nu\cos^2\alpha\right)}{4\pi(1-\nu)} \ln\left(\frac{R_{\mathrm{ext}}}{r_0}\right). \end{aligned} \tag{6.2}$$

Here, R_{ext} is the outer radius (distance from the dislocation core) and r_0 is the core radius, which take the values $R_{\mathrm{ext}}/r_0 = 4$ and $G = 84.85\,\mathrm{GPa}$ in our calculations.

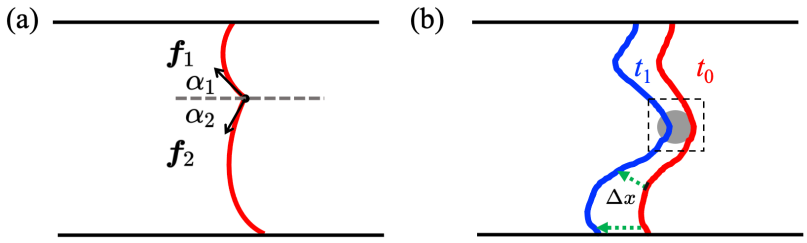

Figure 6.1. Schematic of the dislocation geometry around a local pinning point. (a) Simplified model of a dislocation pinned at an obstacle. (b) Two different configurations of the dislocation at different times. During the time interval Δt, the dislocation segment travels a distance Δx.

As shown in Figure 6.1a, the dislocation is pinned by an obstacle. Close to the pinning point, the force vectors arising from line tension are denoted by $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, respectively. The dislocation line tangents form the angles α_1 and α_2 with the

horizontal direction (indicated by the dashed line). We assume that the pinning point exhibits an interaction strength only along the horizontal direction. Then, based on the force balance, we have

As shown in Figure 6.1a, the dislocation is pinned by an obstacle. Close to the pinning point, the force vectors originating from line tension are $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, respectively. The dislocation line tangents form the angles α_1 and α_2 with the horizontal direction (indicated by the dashed line). We assume that the pinning point exhibits an interaction strength only along the horizontal direction. Then, based on the force balance, we have

$$\|\boldsymbol{f}_{\text{obs}}\| = \|\boldsymbol{f}_1\| \cos\alpha_1 + \|\boldsymbol{f}_2\| \cos\alpha_2. \tag{6.3}$$

6.3. Results and Discussion

This section presents a data-driven analysis of the dislocation dynamics in the investigated HEA, based on *in situ* TEM observations. The workflow begins with the three-dimensional reconstruction of the dislocation microstructure from the recorded TEM video frames, enabling the quantification of individual dislocation geometries and their evolution. Subsequently, local pinning points are identified and analyzed in terms of their spatial location and temporal evolution of strength. Finally, a spatio-temporal coarse-graining approach is applied to the dislocation fields, allowing a quantitative "scanning" of the entire slip plane. This framework treats dislocations as local probes of the atomic-scale energy landscape, providing previously inaccessible information about the local pinning and its dynamic evolution in the representative HEA.

6.3.1. 3D Reconstruction of the Dislocation Structure

A prerequisite for reconstructing the three-dimensional (3D) pile-up structure from TEM images is the precise knowledge of the Euler angles and the tilt convention used during the experiment. These parameters are the key input, apart from the images themselves, for the subsequent 3D reconstruction. The detailed workflow, referred to as the "3D reconstruction technique", is summarized in Section 3.3.2. Based on this framework, the dislocations observed in the TEM projections can be mapped back to their actual spatial arrangement along the corresponding slip plane.

In the present study, the analyzed TEM video, in which several dislocations are repeatedly pinned, has a total duration of approximately four minutes. During preprocessing, a clearly visible reference point, marked as ⊡ in Figure 6.2a, was used to stabilize the video [170]. The stabilized sequence was then accelerated to a playback length of 58 seconds while maintaining a frame rate of 30 fps, yielding 1743 frames in total. Unless otherwise stated, all discussions below refer to the real-time scale of the original four-minute video. The open-source software *labelme* [171] was employed to manually annotate each dislocation in the pile-up by tracing thin lines along their contours. This process generated polygonal datasets whose vertex coordinates represent the dislocation geometry. The dislocations were indexed from left to right,

and for each dislocation, the polygon vertices were indexed from bottom to top. For the entire 58-second video, at least one frame per second was labeled, resulting in a total of 303 annotated frames, each containing the complete dislocation configuration of that moment.

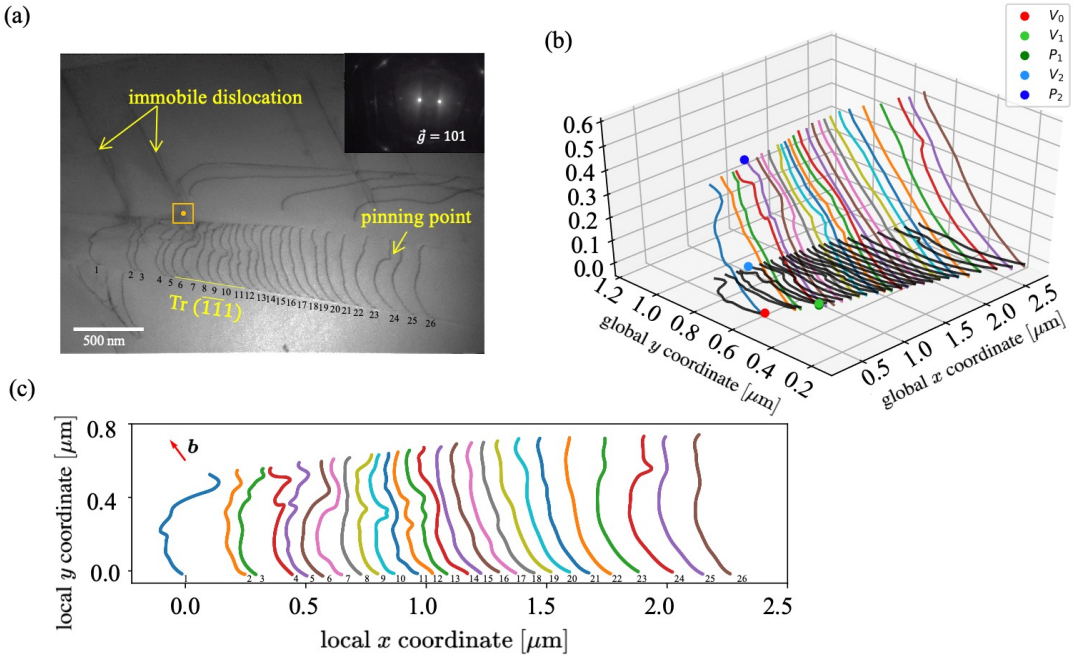

Figure 6.2. Frame #0561 (t = 75 s) showing piled-up dislocations in a CoCrFeMnNi high-entropy alloy during *in situ* TEM straining at 96 K. (a) Pile-up in the image plane. The image was kindly provided by Dr. Marc Legros within the European Project MuDiLingo. (b) Configuration of the pile-up on the image plane with the corresponding diffraction pattern shown in the upper right. (c) Reconstructed configuration of the pile-up on the $[\bar{1}\,\bar{1}\,1]$ glide plane. The red arrow indicates the Burgers vector direction [−1, 1, 0].

Using the experimentally determined parameters summarized in Table 6.1, the 3D dislocation microstructure was reconstructed from the TEM projection (Figure 6.2a). The resulting 3D configuration is presented in Figure 6.2b and further projected onto the slip plane in Figure 6.2c. The reconstruction reveals that the dislocations exhibit varying "heights", corresponding to differences in the z-direction between the dislocation endpoints. Since a single dislocation that does not intersect other crystallographic defects can only terminate at the specimen surface, the variation in height reflects a change in local specimen thickness—an effect typically observed in electrochemically polished samples. In the present case, the reconstructed sample thickness ranges from 460 nm to 560 nm, which is relatively large for thin-foil TEM specimens. This range was independently verified by cross-sectional focused ion beam (FIB) inspection.

The shape of a dislocation is influenced by the interactions with other dislocations, the free-surface effect, the external loading, and local pinning effect. The first two

Table 6.1. Definition and values of the symbols used in the reconstruction. Subscripts "c" and "w" refer to the crystal and world coordinate systems (CCS and WCS), respectively

Symbol	Definition	Value
(ϕ_1, Φ, ϕ_2)	Bunge Euler angles	$(69.3°, 39.5°, -68.1°)$
$\boldsymbol{R}$	Incident electron beam direction (reversed)	(0, 0, 1)
$\boldsymbol{n}_\mathrm{c}$	Slip plane normal in CCS	$[\bar{1}, \bar{1}, 1]$
$\boldsymbol{n}_\mathrm{w}$	Slip plane normal in WCS	[−0.125, −0.745, 0.655]
$\boldsymbol{b}_\mathrm{c}$	Burgers vector in CCS	(−1, 1, 0)
$\boldsymbol{b}_\mathrm{w}$	Burgers vector in WCS	[−0.499, 0.618, 0.607]
θ	Tilt angle (right-hand rule convention)	2.5°

contributions do not vary much when we study an isolated dislocation at the tail of the pile up. The resolved shear stress on a dislocation is estimated through the simple dislocation circle model $\tau = \mu b/R$ where R is the radius of the fitted dislocation circle and μ is the shear modulus (84.85 GPa [59]) and b is the magnitude (0.254 nm) of the Burgers vector. For dislocations at the tail of the pile up, the estimated resolved shear stress is 27.57 ± 3.81 MPa, where the error margin is the standard deviation of estimations over time. Given that the variation is relatively small, we therefore exclude the effect of the external stress on the evolution of dislocation shapes, i.e., change of the dislocation shape mainly results from the local pinning points.

6.3.2. Identification of the Local Pinning Point and Evolution of Strength

Unlike inclusions or second-phase particles, local pinning points are not directly visible in TEM images. Nevertheless, their presence can be inferred from their influence on dislocation motion. In particular, a pinning point locally reduces dislocation mobility, leading to a characteristic bowing of the dislocation opposite to the glide direction, as shown in Figure 6.2a. Consequently, the dislocation itself can serve as a probe for identifying local pinning points, where regions of high curvature along the dislocation line are potential locations of such pinning points.

To identify these local pinning points, three candidate sites, denoted as P_1, P_2, and P_3, were manually selected from the TEM video. At these positions, dislocations were consistently observed to bend sharply while passing by. For each pinning point, the curvature of the dislocation line was quantified within the effective zone of the pinning point, indicated by the dashed square in Figure 6.1b. It should be noted that the maximum curvature of a dislocation sometimes occurs near its endpoints, which is primarily due to surface-related friction effects arising from the surface oxides and image forces at the free surface. The true pinning region corresponds to the local maximum curvature located within the effective zone rather than at the specimen boundaries. In total, 19, 11, and 50 configurations were recorded for P_1, P_2, and

P_3, respectively. The corresponding dislocation indices and the number of captured configurations are summarized in Table 6.2. For instance, at P_1, dislocation ID 15 was pinned twice, ID 16 nine times, and ID 17 eight times during the experiment.

Table 6.2. Summary of dislocation indices and the number of recorded configurations corresponding to the three candidate pinning points P_1, P_2, and P_3

Pinning point P_1		**Pinning point** P_2		**Pinning point** P_3	
ID	No. of configs	ID	No. of configs	ID	No. of configs
15	2	18	6	22	2
16	9	19	1	23	39
17	8	20	4	25	4
				26	5

The strength of each pinning point was estimated using the line tension model illustrated in Figure 6.1a. According to this model, the bending angle of a pinned dislocation is inversely related to the local pinning strength; a smaller bending angle implies a stronger pinning point. As established in the previous section, the influence of external stress on dislocation shape evolution can be neglected. However, since the analyzed dislocations belong to a pile-up configuration, dislocation–dislocation interactions may also contribute to the observed curvature. To evaluate this contribution, the average spacing between each pinned dislocation and its two nearest neighbors (front and back) was measured.

As shown in Figure 6.3a, the average neighbor distance remains nearly constant for dislocations pinned at P_1 and P_2, suggesting that interaction effects are negligible. Hence, the evolution of the bending angle primarily reflects changes in the intrinsic pinning strength rather than in the local stress environment. The corresponding temporal evolution of the pinning strength, derived from the measured bending angles, is displayed in Figure 6.3b. For P_1 and P_2, the pinning strength slightly increases over time as successive dislocations pass by. For example, the strength estimated from dislocation ID 16 is higher than that from ID 15, which traversed P_1 earlier. In contrast, the pinning strength at P_3 exhibits a gradual decrease, particularly for dislocations ID 22, ID 23, and ID 25, where the spacing to neighboring dislocations remains unchanged.

This finding highlights the difference in the strengthening mechanism between HEA and conventional solute-hardened alloys: In solute-hardened alloys, the solid solution hardening comes from the interactions between dislocations and the solution stress field (together with the solution strength when a dislocation cuts through the solute). The effective strength of these solutes is generally constant or decreases after being sheared. However, in the studied HEA, the effective strength of the pinning points (lattice distortion from local ordering) evolves in a 'random' manner. Some pinning points are hardened while others are weakened. The strength of the pinning point

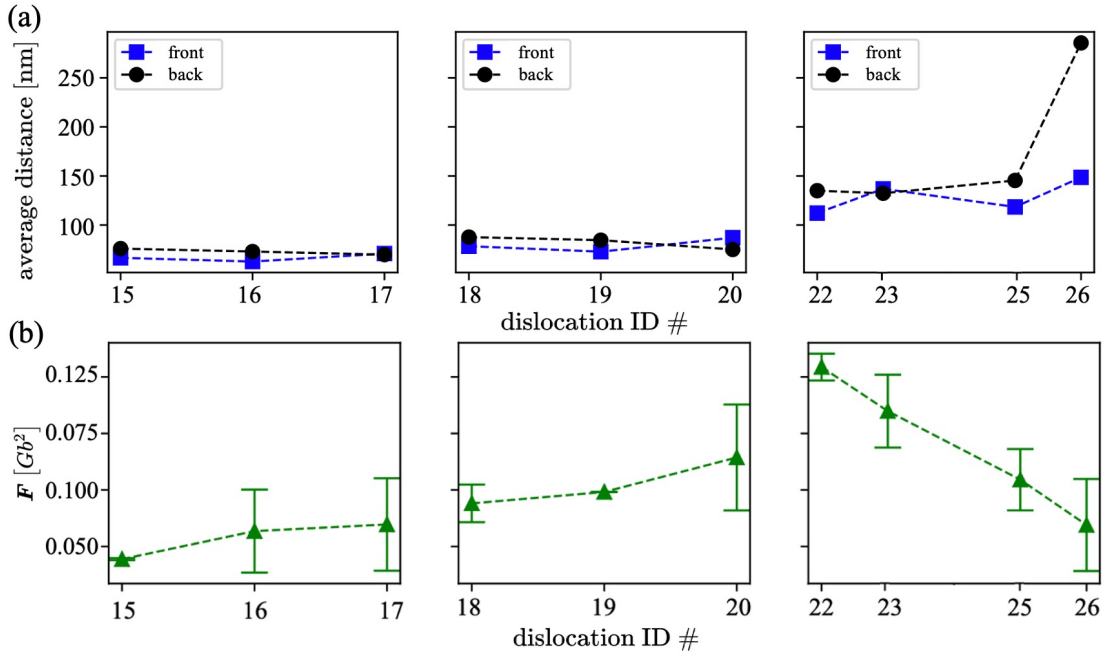

Figure 6.3. Dislocation pinning behavior over time. (a) Variation of the distance between a pinned dislocation and its nearest neighbors (front and back). (b) Evolution of the local pinning strength estimated from multiple passing dislocations. The approximate locations of the analyzed pinning points are marked by cyan boxes in Figure 6.4b. From left to right, the points correspond to P_1, P_2, and P_3, with respective observation intervals of 1.3 s to 14.2 s, 14.4 s to 24.5 s, and 0 s to 23.6 s.

changes all the time in the process of dislocation gliding.

6.3.3. Spatio-Temporal Analysis of the Local Pinning Points through Coarse Graining of Dislocation Microstructure

So far, local pinning points have been identified visually, and their strength evolution was inferred from the variation in the dislocation bending angle during successive passings. In this section, a spatio-temporal coarse-graining analysis of the dislocation microstructure is performed, which allows the simultaneous visualization of these local pinning points. The underlying concept is that dislocations can be regarded as dynamic probes sampling the local atomic configuration of the material. In response to local structural fluctuations, dislocations deform under an external stress field, which is reflected in the local curvature of their line shape and in variations of their local velocity. These two quantities, curvature and velocity, represent the measurable "signals" through which the local microstructural environment can be inferred. Conversely, when dislocations interact with obstacles, they also modify the local atomic configuration, thus showing a "action–reaction" process.

By scanning the entire slip plane with dislocations as local probes and evaluating the two signals mentioned above, a spatially resolved representation of the underlying pinning landscape can be obtained. As in any scanning-based method, the resolution is a crucial parameter. Here, the resolution is defined by the voxel or pixel size used to discretize the slip plane. Within each pixel, the signals (curvature and velocity) are averaged spatially and temporally. The spatial averaging of all considered dislocations yields the coarse-grained dislocation fields, an effective representations of the complex dislocation microstructure [4]. The temporal averaging corresponds to an ensemble average over all scanned frames; however, it is not a simple arithmetic mean. Instead, the contribution of each voxel is weighted by the local coarse-grained dislocation density, meaning that only voxels containing dislocation segments contribute to the average. A detailed description of the spatio-temporal coarse-graining scheme and its numerical implementation is provided in Section 4.2.3.

An example of such a coarse-grained slip-plane analysis is shown in Figure 6.4, where a coarse-graining resolution of 15 nm was employed. Reducing the pixel size further did not yield notable differences. The dislocations identified in the TEM images (Figure 6.4a) were first reconstructed in 3D and subsequently projected onto their corresponding slip plane. The coarse-grained curvature field obtained from this configuration is shown in Figure 6.4b. The curvature varies significantly along individual dislocation lines, reflecting the characteristic wavy and rough morphology typical of dislocations in HEAs. The coarse-grained curvature assumes both positive and negative values depending on the direction of bending. In this convention, curvature is defined as positive when the dislocation bows opposite to the glide direction (i.e., toward the positive x-direction). The three previously identified pinning points are indicated in the lower panel of Figure 6.4, showing strong positive curvature values around their locations.

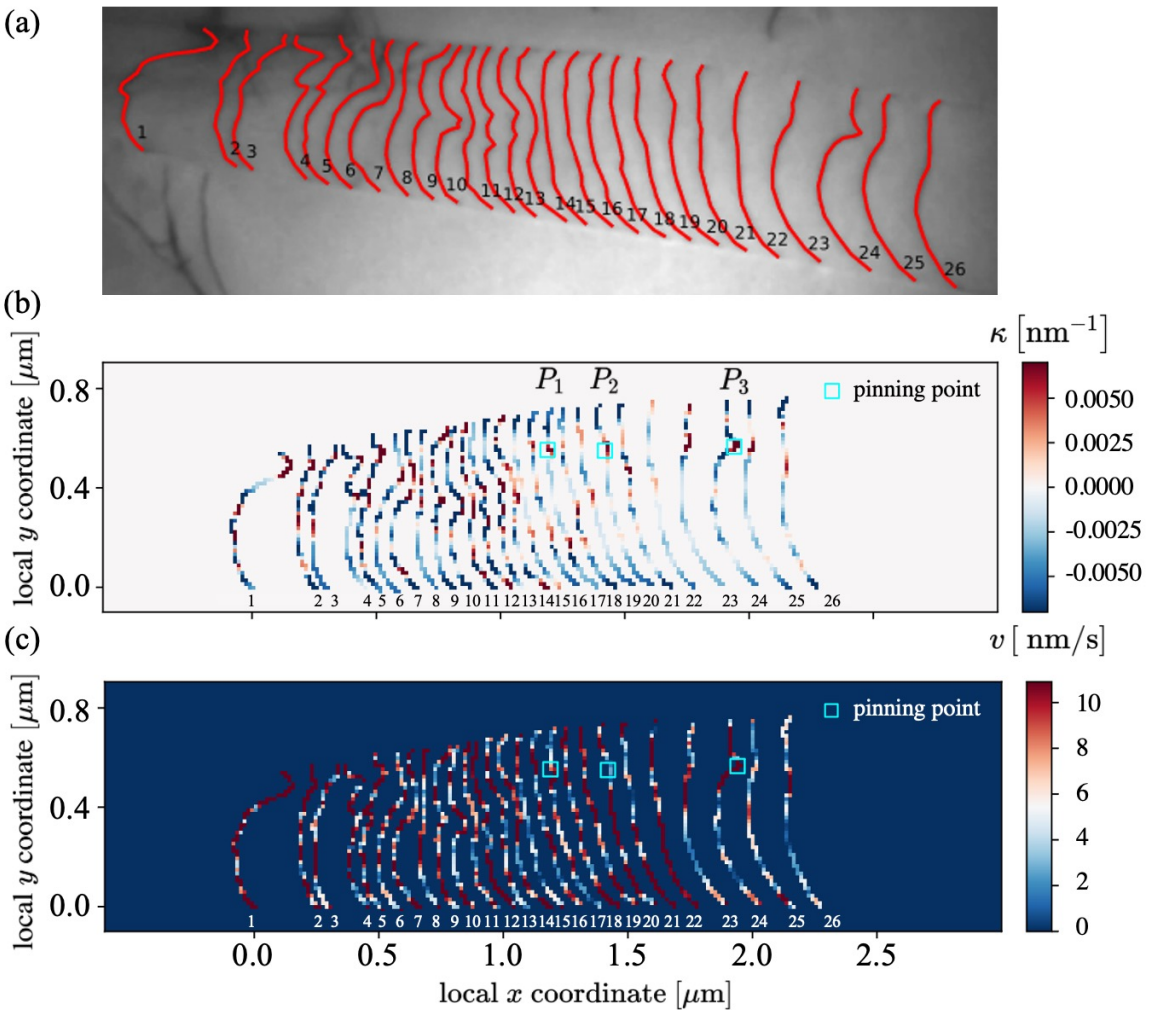

Figure 6.4. Example of coarse-graining analysis using dislocations as local probes. (a) Labeled dislocations in a frame from the original *in situ* TEM video, acquired by Dr. Marc Legros within the European Project MuDiLingo. (b) Coarse-grained dislocation curvature, computed with a pixel size of 15 nm. (c) Coarse-grained dislocation velocity calculated from two consecutive frames.

The procedure for determining the local dislocation velocity is schematically illustrated in Figure 6.1b. Based on the polygonal representation of each dislocation, the velocity of a segment is calculated as $\Delta x/\Delta t$, where Δx is the perpendicular distance traveled by the segment during the time interval Δt. The black line in Figure 6.1b marks the initial dislocation position. The time interval between two frames (t_0 and t_1) is chosen sufficiently small to ensure accurate displacement measurement. For the endpoints of each polygon, Δx is directly computed as the distance between the corresponding end positions, indicated by the green dashed line in Figure 6.1b. The resulting coarse-grained velocity field, calculated between two consecutive frames, is displayed in Figure 6.4c. It reveals significant velocity variations along the dislocation line. Notably, near P_3, the velocity on the left (already depinned) side is higher than that on the right (still pinned) side. This behavior is consistent with the classical picture of dislocation depinning in a quenched disorder landscape: in the pinning regime, the velocity of the dislocation is slowed down while in the depinning regime, the sudden increase in velocity leads to the dislocation avalanche.

The spatio-temporal coarse-graining over all 303 frames yields the distributions of curvature and velocity shown in Figure 6.5a and Figure 6.5c, respectively. In Figure 6.5a, several bands of high curvature appear near the head of the pile-up (left region of the slip plane), which correspond to low-velocity bands in Figure 6.5c. These correlations mark regions of intense local pinning activity. Additional high-curvature bands, highlighted in magnified regions, are also observed farther away from the main pile-up head. Interestingly, these bands are approximately aligned with the Burgers vector direction. Among the highlighted regions, the rightmost band coincides with the previously analyzed pinning point P_3. The formation of this band can be explained in two ways: (i) the presence of multiple adjacent pinning points, or (ii) the gradual translation of a single pinning site due to its interaction with passing dislocations. The latter explanation appears more plausible, as pinning points in HEAs originate from local atomic ordering. When a dislocation glides, the relative shift of atomic planes by one Burgers vector alters this local ordering, effectively dragging the pinning site along the slip direction. This behavior differs from both classical solute hardening—where solutes are assumed immobile—and from the Cottrell drag mechanism, in which mobile solutes follow moving dislocations. It also rules out pinning by immobile jogs, whose positions would remain unchanged.

The fine-scale coarse-graining captures local variations in dislocation behavior, while additional averaging along the y-direction, as shown in Figure 6.5b and Figure 6.5d, reveals how curvature and velocity evolve along the pile-up direction (x-axis). Our probe for sampling the slip plane is a dislocation pile-up structure (even though the particular structure is not ideal since the leading dislocation is not entirely blocked). For an ideal dislocation pile-up, it is expected that the velocity of dislocations decreases from the tail to the head of the pileup. Any deviation from this trend would result from the local pinning effect. Indeed, Figure 6.5b shows numerous high-curvature regions along the pile-up, reflecting the widespread presence of pinning effect. The resulting velocity profile (Figure 6.5d) differs markedly from the idealized behavior. In particular, region II exhibits an almost constant velocity despite pronounced

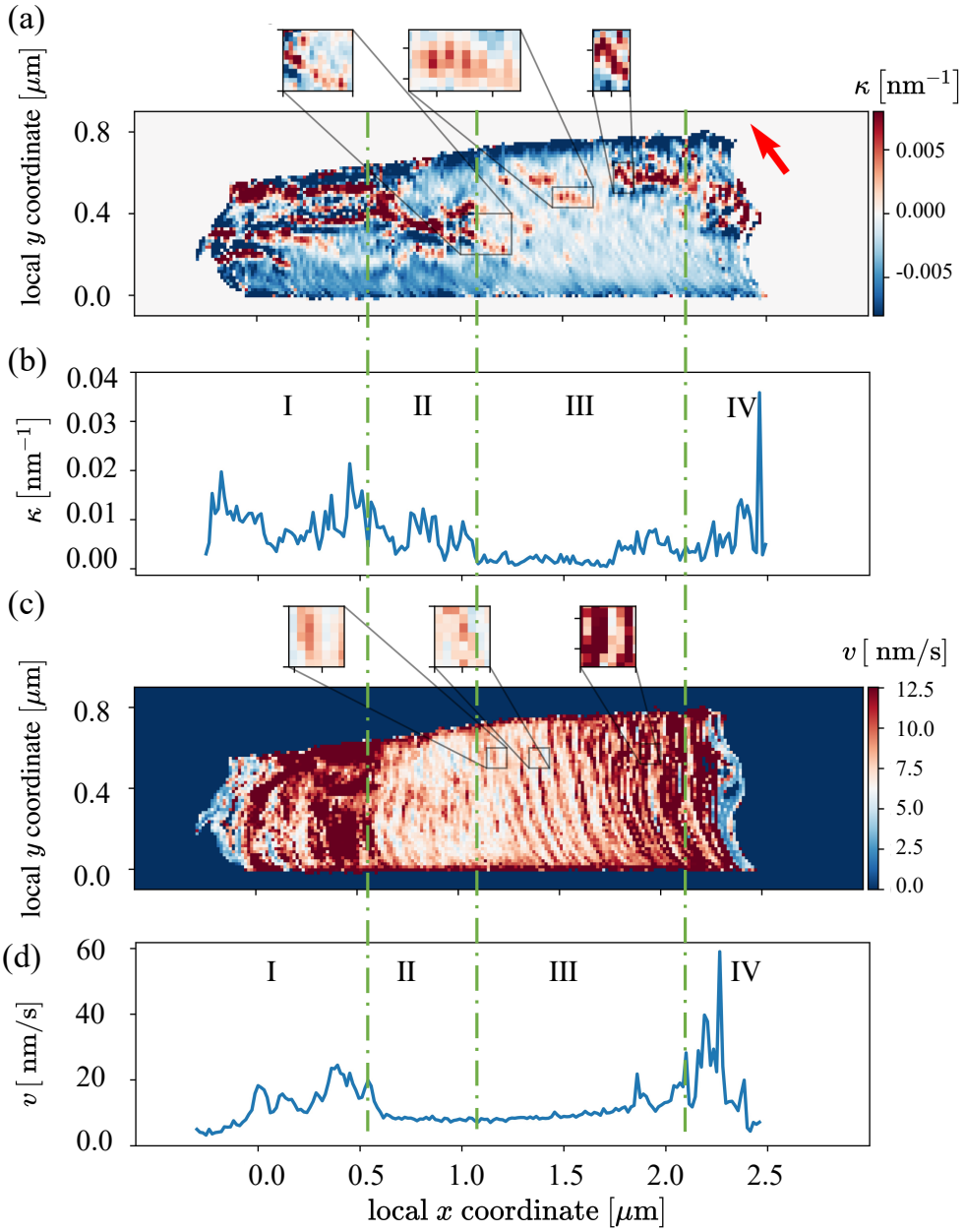

Figure 6.5. Spatio-temporal averaging of curvature and velocity over 303 frames. (a) Distribution of curvature with a pixel size of 15 nm. The red arrow indicates the Burgers vector direction $[-1, 1, 0]$. (b) Averaged curvature (along y-direction) plotted along the pile-up direction (x-axis). (c) Distribution of velocity with a pixel size of 15 nm. (d) Averaged velocity (along y-direction) plotted along the pile-up direction (x-axis).

curvature fluctuations, implying a dense cluster of pinning points where dislocations are repeatedly trapped and released over short distances. Isolated pinning–depinning events, evident from sharp velocity fluctuations (e.g., to the right of region III and left of region IV in Figure 6.5c), further support this interpretation. In principle, our method can also identify each single pining point in such dense regions, however, this would require a much better resolution both spatially and temporally (limited by the experimental video frame rate).

6.4. Conclusion

In this chapter, we investigated the pinning point dynamics of the equimolar CoCrFeMnNi HEA at near liquid nitrogen temperature using in situ TEM straining. For this, we introduced a new data mining approach that relies on spatio-temporal coarse graining of pile-up dislocations used as probes of the local energy landscape. Such analysis of the morphology and motion of dislocations leads to the following conclusions:

1. The strength of the pinning obstacles changes upon dislocations passing by. The change in strength is random as some obstacles get weakened while others get strengthened.

2. The spatial location of the pinning obstacles is observed to vary along the Burgers vector direction.

3. The spatial distribution of the pinning obstacles is not homogeneous: there are isolated pinning obstacles and close-packed pinning obstacles.

4. In the specific case of HEAs, this approach establishes for the first time that moving dislocations experience a friction arising from obstacles that have random strength and distribution and that this randomness is shifted in time and space by passing dislocations; the friction effect is not completely annihilated.

Our data mining approach—using dislocations as probe, and the coarse-grained field quantities as resulting 'scanning' signals—is able to detect the pinning point locations and strength. It is thus a very promising technique for further fundamental *in situ* TEM studies of dislocation microstructures and the related property of materials.

7. Unsupervised Learning of Nanoindentation Data: Inferring Microstructural Details of Cu–Cr Composites

7.1. Introduction

Building on the probabilistic framework introduced in Chapter 5, this chapter explores its use in nanoindentation data interpretation. In this study, we utilized Cu–Cr composites with controlled heterogeneity as a model material, characterized by varying fractions of the two material components. Our objective was to assess the effectiveness of the statistical nanoindentation technique in detecting variations in mechanical properties within the material. While the microstructures of the selected materials allow for precise positioning of indents and a conventional analysis, this controlled heterogeneity served as a basis for exploring fundamental methodological aspects, such as data processing, uncertainty quantification, and model selection. Unlike more complex materials with multiple, possibly unknown, phases and heterogeneous microstructures, the controlled heterogeneity in our composites simplifies the interpretation of the nanoindentation data by eliminating the complexities associated with unknown heterogeneity. Another key question that arises in the context of statistical nanoindentation is: How much data is necessary to obtain reliable and meaningful results? We address this question by systematically investigating the effects of data quantity on the statistical analysis of the nanoindentation data. The content of this chapter has been published in [172].

7.2. Materials and Methodology

Cu–Cr composites as a two-phase model material with different compositions were evaluated. Four materials were studied with different fractions of Cr, i.e., 25 wt% Cr and 60 wt% Cr corresponding to 29.95 at% and 64.40 at% Cr, respectively, as well as Cu and Cr as reference samples. All materials were produced via field-assisted sintering technique (FAST) as described in detail in [173]. Briefly, Cu powder with 99.9 at% purity and technically pure Cr powder (99.5 at%) were used and compacted at a temperature of 950 °C and a pressure of 40 MPa, except for the Cr sample that

was compacted at a temperature of 1450 °C. The composite samples (in Figure 7.1) will be referred to as CuCr25 and CuCr60 according to their nominal compositions. For indentation testing, the sample surfaces were prepared applying standard grinding and polishing techniques using SiC paper and diamond suspensions with decreasing grain size down to 0.1 µm.

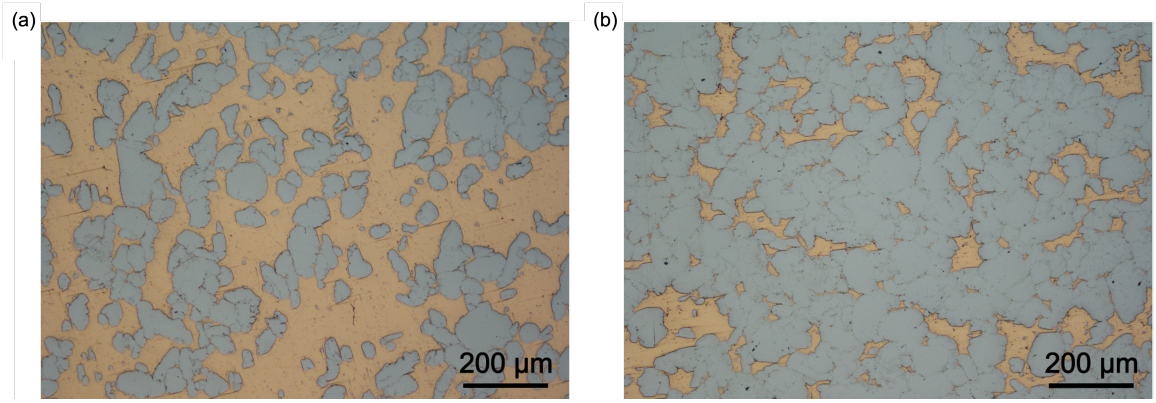

Figure 7.1. Optical micrographs of Cu–Cr composites produced by field-assisted sintering technique with (a) 25 wt% Cr (referred to as CuCr25). (b) 60 wt% Cr (referred to as CuCr60). The samples and micrographs were kindly provided by Prof. Dr. Ruth Schwaiger and the research group, and were investigated by indentation.

A nanoindenter G200 XP (Agilent/Keysight Technologies, Inc., CA, United States) equipped with a diamond Berkovich tip was used to investigate the mechanical properties of the composite samples. The samples were indented to different depths using the so-called Express Test option. Arrays of indents covering areas up to 500 µm × 500 µm were made in different locations on the sample surface. The indentation depths ranged from 200 nm to 2000 nm and the distance between individual indents was maintained approximately constant between 20 µm and 23 µm for all depths. Hardness H and Young's modulus E were determined assuming 1141 GPa and 0.07 for Young's modulus and Poisson's ratio, respectively, of the diamond tip. The Poisson ratio of the materials was fixed at 0.25. Figure 7.2 shows an example distribution of hardness and Young's modulus of CuCr60 for an average indentation depth of 1 µm. The marginal hardness distribution is a bimodal distribution, while a trimodal distribution can be seen for the marginal histogram of Young's modulus.

The Cu–Cr composites consisted of two distinct phases with average Cr particle diameters of approximately 30 µm on the indented surface; thus, upon indenting the surface we assume that either one or the other element dominates or the properties of a mixture of both phases is measured. However, the "mixture of elements" might as well be more than "just the sum of its parts" since additional effects, e.g., related to the presence of interfaces, might occur during indentation. Furthermore, we assume that similar local microstructural or chemical properties lead to similar measurement data and that the mechanical properties exhibit gradual changes over the surface.

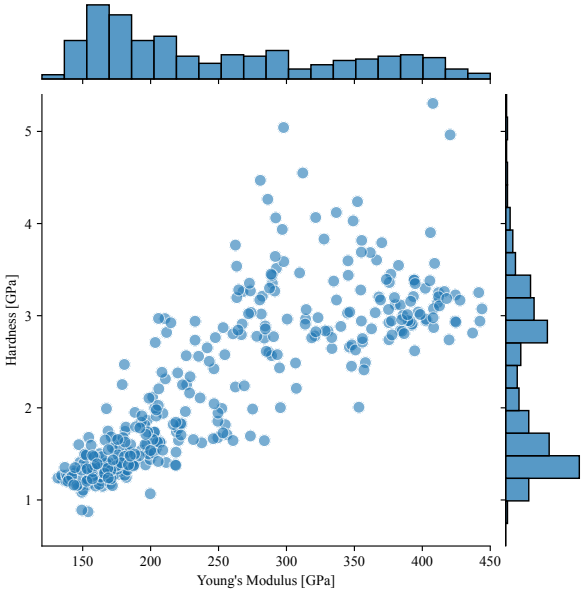

Figure 7.2. Scatter plot with marginal histograms of all obtained values for Young's modulus and hardness measured for CuCr60 at an average indentation depth of 1 µm.

Thus, the data-science task is to analyze a number of data records consisting of the given feature variables E (Young's modulus) and H (hardness), measured for four different materials and at different depths, i.e., 300 nm, 400 nm and 1000 nm. The whole dataset $\mathcal{D}$ consisting of N data records, can be written as the set of pairs

$$\mathcal{D} = \{(E_i, H_i)\}_{i=1\ldots N} \,. \tag{7.1}$$

The goal of this work is to perform a clustering analysis, during which data points with similar elastic properties are grouped together, i.e., for each pair (E_i, H_i) in the feature space, we determine the categorical variable y_i that gives the number of the respective cluster. Additionally, the total number of different cluster needs to be determined. Since annotated training data does not exist, this has do be done in an unsupervised manner.

For the clustering, we have chosen the GMM, which is one of the most-studied probabilistic clustering models in the statistical literature [21, 174, 175]. Using nanoindentation data to investigate the distribution of heterogeneous materials has great potential, provided the GMM technique is applied properly and a sufficient amount of data is available. Here, each "phase" is assumed to correspond to an individual Gaussian distribution of Young's modulus and hardness. These materials "phase" are commonly referred to as *components* in the context of machine learning. We assume that the distribution of experimental data was generated by a combination of Gaussian processes, which are represented by the probability density functions (PDFs) $\mathcal{N}$ for

each component j. The resulting superposition is then given by

$$p(\mathcal{D};\boldsymbol{\phi}) = \sum_{j=1}^{k} \alpha_j \mathcal{N}(\mathcal{D} \mid \boldsymbol{\theta}_j). \tag{7.2}$$

Here, k is the number of components of the model, $\alpha_j > 0$ are the weights of each component j, the $\boldsymbol{\theta}_j$ are the vectors of parameters for the Gaussian, and $\boldsymbol{\phi} = \{\alpha_1, ..., \alpha_k, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_k\}$ is a short notation for the whole set of parameters governing the Gaussian mixture model.

For a multivariate Gaussian, the component j of the superimposed function is given by the parameters $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the mean vector and the covariance matrix, respectively. $\boldsymbol{\mu}_j$ describes the location of the component j in the feature space, while the covariance matrix $\boldsymbol{\Sigma}_j$ characterizes the j-th component data distributed around $\boldsymbol{\mu}_j$. The objective of the training process is to estimate the values for the model parameters of the Gaussian distribution that best align with the training data $\mathcal{D}_{\text{train}} \subset \mathcal{D}$. The model parameters $\boldsymbol{\phi}_k$ are iteratively determined while assuming the number of clusters k to be predefined. Typically, the superposition of Gaussians is computed using the maximum likelihood method based on a set of candidate models that differ in the number of clusters generated using the expectation maximization algorithm.

The Bayesian information criterion (BIC) is employed as a selection criterion to identify the optimal model. The model with the lowest BIC value indicates the highest likelihood which has the best predictive capability for the observed data. In this study, the GMM analysis has been implemented using the open-source Python package "scikit-learn" [176]. A comparison in Section 7.3.4 additionally justifies the choice of the GMM model.The datasets generated and analyzed during the current study are available in the online repository. Further details of the algorithm employed can be found in the supplementary material.

7.3. Results and Discussion

In the following, we will start with the description of the data cleaning process and the analysis of the one-component metallic composites using a 1D Gaussian mixture model. Then, the CuCr25 and CuCr60 composites are investigated using both 1D and 2D GMM for an average indentation depth of 1 µm. Finally, a comprehensive analysis is conducted to examine the influence of sample size on the robustness of the model as well as the influence of indentation depth.

7.3.1. Preparation of the Datasets

Figure 7.3a illustrates the original datasets comprising the measured E and H values of the four materials tested. Due to variations in the height of the materials' regions on the polished surface, the actual recorded maximum indentation depths ranged from 200 nm to 2000 nm, deviating from the nominal values set in the experiments. Here, we

focused our analysis on depths ranging from 800 nm to 1200 nm. Subsequently, the data underwent cleaning and filtering processes to remove measurement errors, particularly outliers with unrealistically high values as well as other invalid data. For CuCr25, data within the ranges of $100 \leq E \leq 400$ GPa and $0.8 \leq H \leq 4.5$ GPa were retained, while for CuCr60 the data range was $100 \leq E \leq 500$ GPa and $1.0 \leq H \leq 5.0$ GPa. The resulting cleaned data, comprising approximately 98% of the original data, is presented in Figure 7.3b. Notably, the distributions of the two pure metals exhibit lower variances compared to the Cu–Cr composites. Further preprocessing of the data, such as standardization, was found to have no significant impact on the training results and was therefore not included in this study.

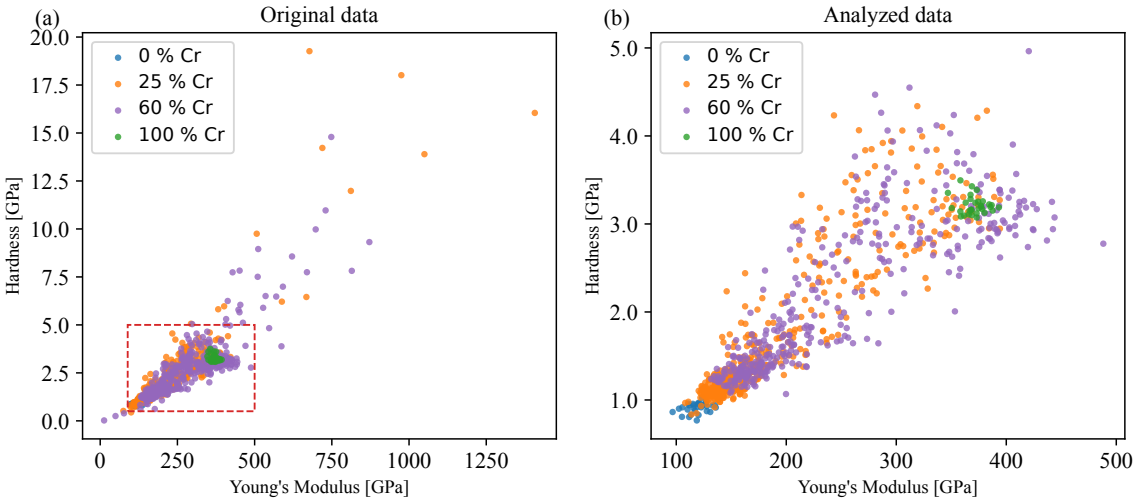

Figure 7.3. Distributions of Young's modulus and hardness obtained experimentally for samples with 0 wt%, 25 wt%, 60 wt%, and 100 wt% Cr content. (a) Raw data as measured. (b) Cleaned and preprocessed dataset employed in the analysis, corresponding to the region highlighted by the rectangle in (a).

7.3.2. Mechanical Properties of the Pure Cu and Pure Cr Specimens

The mechanical properties of pure Cu and pure Cr were analyzed for the indentation depth of 800 nm ~ 1200 nm. The probability density function (PDF) plots of E and H are shown in Figure 7.4a–e. These plots illustrate the mean values of E = 118.80 GPa and H = 0.91 GPa for pure Cu, and E = 371.24 GPa and H = 3.21 GPa for pure Cr. The bin size was selected to ensure an approximately equal number of bins covering the range of all PDFs. To utilize the GMM, it is important to demonstrate that the distributions are roughly normal distributed, despite the GMM's inherent robustness. Thus, the Shapiro-Wilk test was employed to assess the normality of the distributions [177], resulting in p–values of 0.95 for E and 0.14 for H in pure Cu, and 0.95 for E and 0.14 for H in pure Cr. The test accepts the hypothesis, when the

p–value exceeds 0.05, confirming that the data was sufficiently normally distributed. The optimal number of components was 1 for both the Cu and Cr specimens (see Figure 7.4c and Figure 7.4f), as determined by the BIC analysis.

Regarding determination of the component numbers, the result of 1D GMM for E with high p–value is therefore more reliable than the fit of H. Nonetheless, because H is strongly correlated with local microstructure details, such as phase and grain boundaries or dislocations, the fitting result of H can be used to infer local microstructural characteristics. This can also been seen from the small variation at the beginning of the BIC plot for hardness in Figure 7.4f, where the BIC values for $k = 2$ differ only marginally from the one at $k = 1$ that we had identified as the optimal one.

7.3.3. Mechanical Properties of the CuCr25 and CuCr60 Specimens

The histogram of E and H of the CuCr composites, together with their best-fit models are shown in Figure 7.5a and Figure 7.5d (top row) and Figure 7.5b and Figure 7.5e (middle row). The bottom panel Figure 7.5c and Figure 7.5f shows the BIC as a function of the number of components.

The BIC values for both the modulus and the hardness cover a range of around 500 and ≈ 200, respectively, if $k = 1$ is excluded. Though not obvious, this is an important difference as compared to the previous investigation of pure Cu and Cr, as detailed below.

A fundamental consideration in statistical modeling is the relationship between the number of free parameters and the amount of available data. Taking CuCr60 as an example, there are $N \approx 300$ valid measurements for E determined at an average depth of 1 µm. Assuming an optimal number of clusters of $k = 3$, the number of parameters to be measured in a 1-dimensional analysis are three mean values (μ_1, μ_2, μ_3), three variance values $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$, and two coefficients that determine the relative weights of the three Gaussian (α_1, α_2). Given that the sum of all weights should be 1, which means $\sum_{j=1}^{k} \alpha_j = 1$, and $\alpha_3 = 1 - (\alpha_1 + \alpha_2)$, the number of parameters to be determines is $d = 8$. Thus, as a first estimate we conclude that the amount of data should be more than sufficient in this case (assuming that the variance of the data is small).

Are the differences of the BIC values as a function of k large or small? To answer this question, we need to understand how much variation results from a small change in the dataset. The logarithm of the likelihood function $\mathcal{L}$ calculated for this dataset ranges from −10 to −4. According to Equation (5.9), the first term $-2 \log \mathcal{L}$ therefore varies between $-2 \times (-4) = 8$ and $-2 \times (-10) = 20$, while the second term $d \ln N$ equals $8 \times \ln 300 \approx 45$. Thus, the resulting BIC value is expected to lie between 53 and 65. Even with small variations in the data (e.g., a single outlier among 300 measurements), the BIC is likely to remain within this range, indicating that it effectively captures changes in the dataset. Given this estimated range, a BIC difference larger than ~ 50 between two competing models would already constitute strong evidence that the corresponding number of components k is indeed the more likely one. In this example,

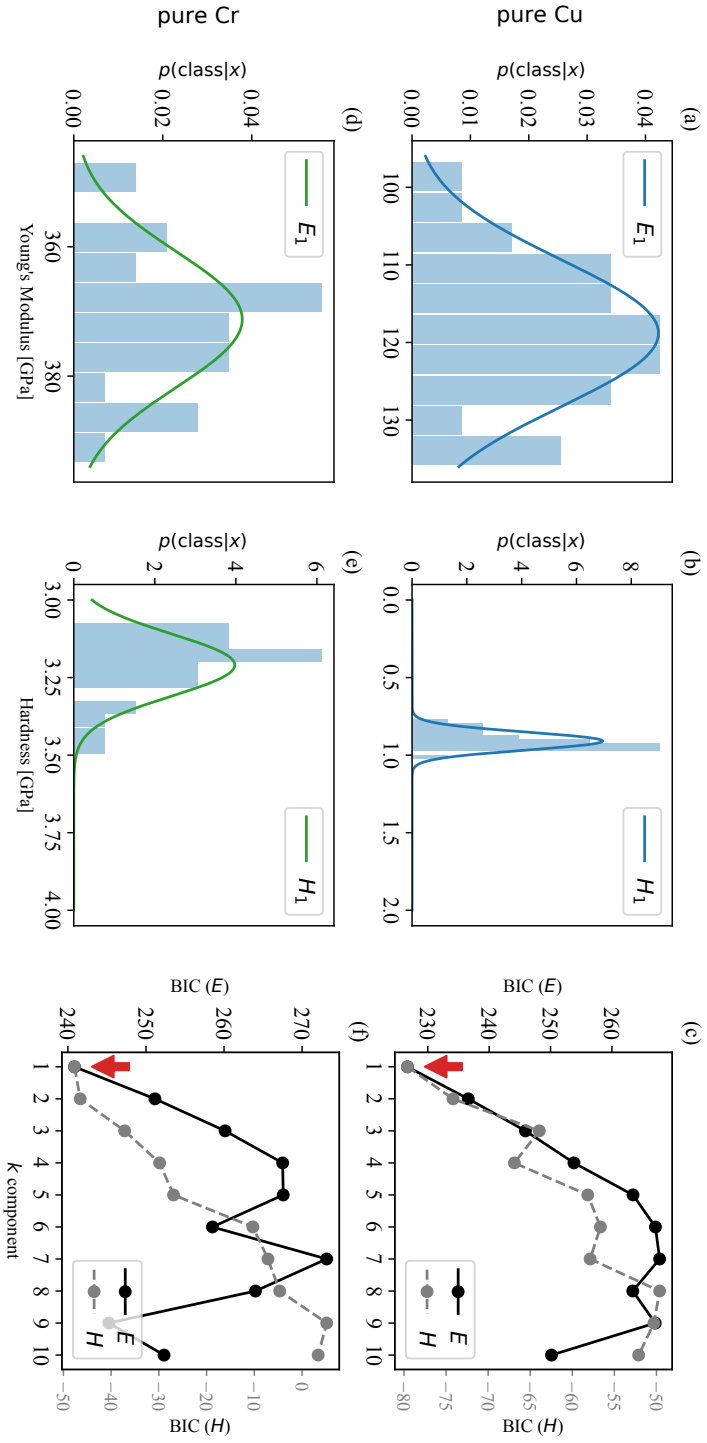

Figure 7.4. Probability density functions (top and middle row) and plot of BIC (bottom row) of pure Cu and Cr. The BIC values are shown for both, Young's modulus (upper y-axis) and hardness (lower y-axis). The left column shows the data for Cr (d–f), while the right column shows those for Cu (a–c).

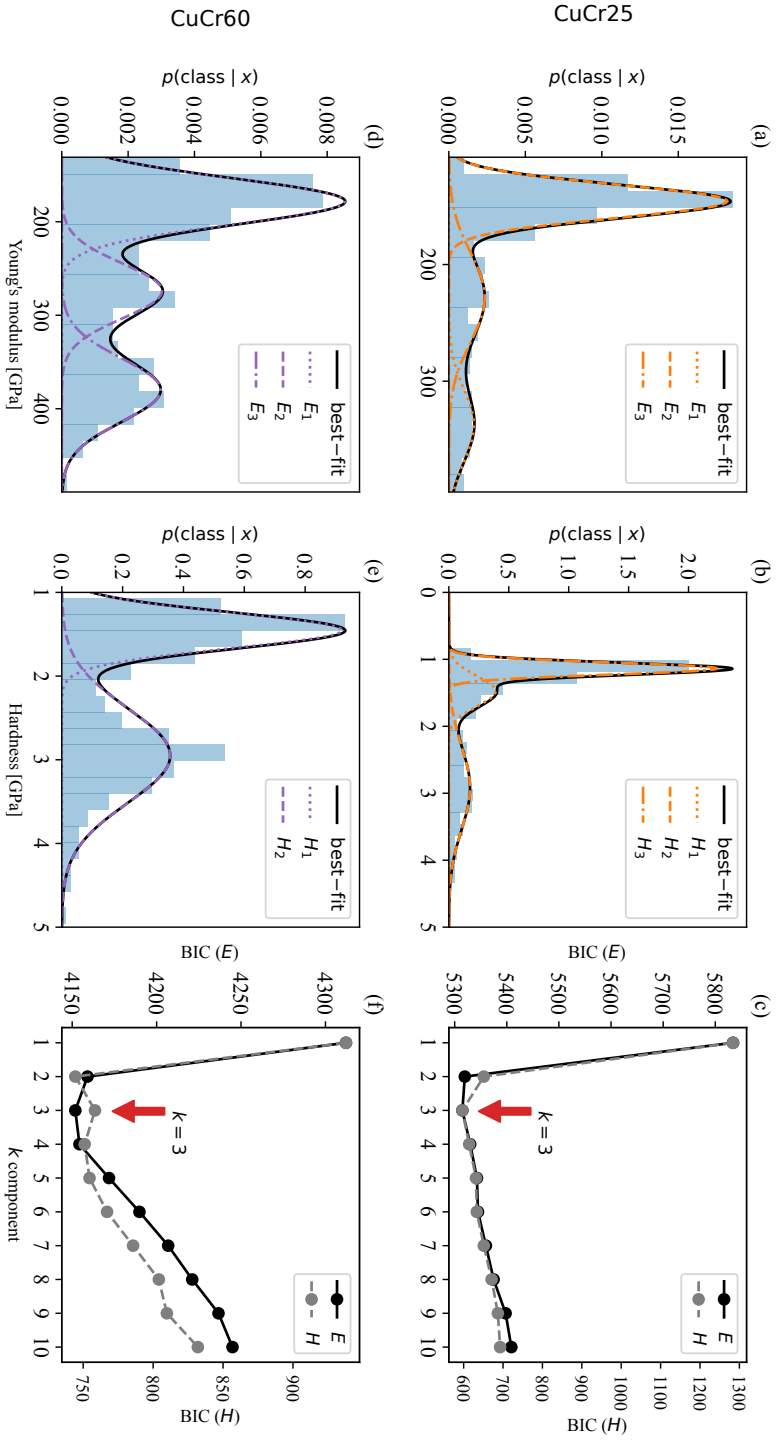

Figure 7.5. 1D GMM results of CuCr25 and CuCr60 at 1 µm depth. CuCr25: (a) Histogram of the E_i and the best fit (solid line). (b) Histogram of the H_i and the best fit (solid line). (c) BIC of H and E; CuCr60: (d) Histogram of E_i and the best fit model. (e) Histogram of H_i and the best fit model. (f) BIC of H and E.

the difference between a BIC value for k and $k+1$ is only slightly smaller which suggests that the differences are relevant and not just "noise". Additionally, taking a look at larger values of k, there is a very clear trend that points to the minimum. In conclusion, our 1D GMM analysis of Young's modulus reveals that both composites have three mechanical phases at 1 µm depth. Table 7.1 summarizes the results of the Gaussian mixture models evaluated for the four different materials.

Table 7.1. 1D mechanical property fitting based on the optimal BIC results at 800 nm ~ 1200 nm depth

Cr (wt%)	No.	Property	Average (GPa)	Std Dev	Pct. (%)
0 (pure Cu)	30	E_1	118.80	9.45	100
		H_1	0.91	0.06	100
25 (CuCr25)	513	E_1	145.55	14.12	64.5
		E_2	226.50	38.42	22.6
		E_3	337.02	31.69	12.9
		H_1	1.14	0.09	50.1
		H_2	1.47	0.23	22.6
		H_3	2.92	0.62	27.3
60 (CuCr60)	364	E_1	177.35	24.17	52.2
		E_2	272.17	29.75	22.4
		E_3	379.32	34.34	25.4
		H_1	1.45	0.22	50.0
		H_2	2.96	0.56	50.0
100 (pure Cr)	31	E_1	371.24	10.54	100
		H_1	3.21	0.10	100

As stated above, our GMM analysis of CuCr25 shows the presence of three mechanical phases as well as differences in the properties of nominally identical phases. In CuCr25, for example, the Cu-rich phase accounts for 64.5 vol% (defined by E_1 as the lowest, closest to pure Cu), while the combined volumes of E_2 and E_3 amount to 35.5 vol%. These results indeed coincide well with the 35.5 vol% Cr estimated by optical microscopy [178]. The lowest modulus value for the Cu phase was found in the compacted Cu sample (i.e., E_1), followed by E_1 of CuCr25 and CuCr60. The highest modulus of the Cr phase was found for CuCr60 (i.e., E_3) and the compacted pure Cr sample (i.e., E_1 in Table 7.1) almost reaching the modulus value of pure Cr.

These differences in the mechanical properties fitted by GMM are likely related to the diffusion of one phase into the other, the presence of foreign particles or pores, or the influence of the surrounding material. Assuming that a small amount of Cr, i.e., 0.4 at% ~ 3 at%, can be dissolved in the Cu matrix [179, 180], the Cr solid solution likely contributes to the difference of the modulus of the Cu phase in the composite samples. In addition, the presence of Cr nanoparticles of 100 nm ~ 200 nm in size was

reported [173], which as well results in a higher Young's modulus value of the Cu phase in CuCr25 and CuCr60. Finally, considering the relative densities of 99.4% and 98.3% for CuCr25 and CuCr60, respectively, compared to 97.7% for the pure Cu sample, a Young's modulus value with an estimated reduction up to 9% [181] can be expected. The reduction of the modulus of the Cr phase in CuCr25 (i.e., E_3) can be explained by the surrounding softer Cu phase. While in general also the hardness fitting of CuCr25 supports the presence of three mechanical phases, the phase fractions are different. Again, this is not surprising, since hardness is a local property and strongly related to microstructural details.

We now take a look at the outcomes of the 2D GMM with independent feature variables E and H shown in Figure 7.6. The distribution of three and four components or mechanical phases of CuCr25 and CuCr60, respectively, are shown in Figure 7.6a-b and Figure 7.6d-e. The red points in the graphs represent the average value of each component, while the concentric ellipses with different orientations (covariance) represent the different components. The model selection criteria BIC are shown as a function of the number of components in Figure 7.6c and Figure 7.5f.

Based on Figure 7.6c, one could say that four mechanical phases are the ideal match for the CuCr25 composite. Given the anomaly in the upper left corner of Figure 7.6b, though, the best assumption remains at three, which will be further discussed in Section 7.3.4. As shown in Table 7.2, in the 2D GMM, which combines both E and H, the estimated amount of Cr in CuCr25 was 38.9 vol%, which is close to the actual experimental findings [178].

Table 7.2. 2D mechanical property fitting based on the optimal BIC results at 800 nm ~ 1200 nm depth

Cr (wt%)	No.	Average E (GPa)	Average H (GPa)	Pct. (%)
25 (CuCr25)	513	144.91	1.17	61.1
		220.92	2.25	27.1
		340.52	3.18	11.8
60 (CuCr60)	364	172.32	1.42	43.4
		262.63	2.67	34.6
		383.35	3.02	22.0

The 2D GMM analysis in Figure 7.5f also reveals that CuCr60 contains three mechanical phases in the depth range of 800 nm ~ 1200 nm. The fitted result for the volume of Cr in 1D is 47.8 vol% based on the modulus values, which is less than the nominal value (i.e., 65 vol% Cr). By contrast, the 2D GMM result indicates a Cr volume fraction of 56.6 vol% (Table 7.2). The difference between the 1D GMM (E) and 2D GMM results is related to H, while the difference between the fitting results and the experimental data is due to the amount of data and variation of microstructures over the samples. Note that the number of datapoints for CuCr60 is by 40% less than that for CuCr25; an insufficient amount of data can be the source of inaccuracies,

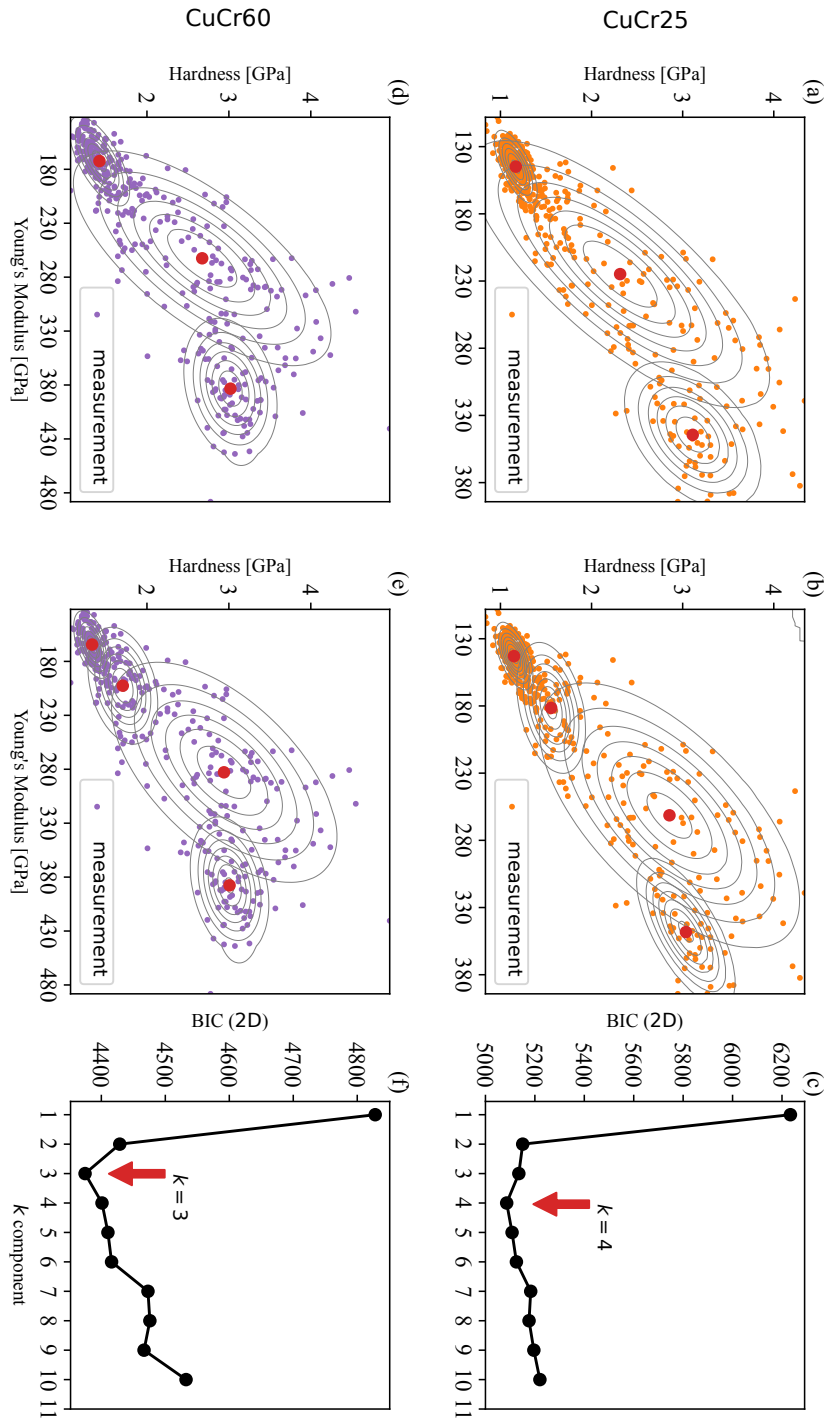

Figure 7.6. 2D Gaussian mixture model clustering of CuCr25 and CuCr60. CuCr25: (a) Three components. (b) Four components. (c) 2D BIC. CuCr60: (d) Three components. (e) Four components. (f) 2D BIC. The ellipses in (a), (b), (d), and (e) are isolines of the Gaussian distributions and the red points represent the average values of the different components.

which we are addressing in the following.

In the above analysis of CuCr60, the size of the dataset was 364, collected over an area of 500 × 500 µm². To increase the size of the dataset, we merged it with two more nanoindentation areas (100 × 100 µm² and 300 × 300 µm²) and analyzed the data using the procedures described above. The results are summarized in Table 7.3 and Table 7.4. Merging the three datasets now includes indentation depths ranging from 500 nm to almost 2000 nm with 97.8% of the data lying between 800 nm–1200 nm indentation depth. As shown in Figure 7.7a, the distribution ranges of the data are congruent indicating that the microstructure of the material was comparable over the different areas indented.

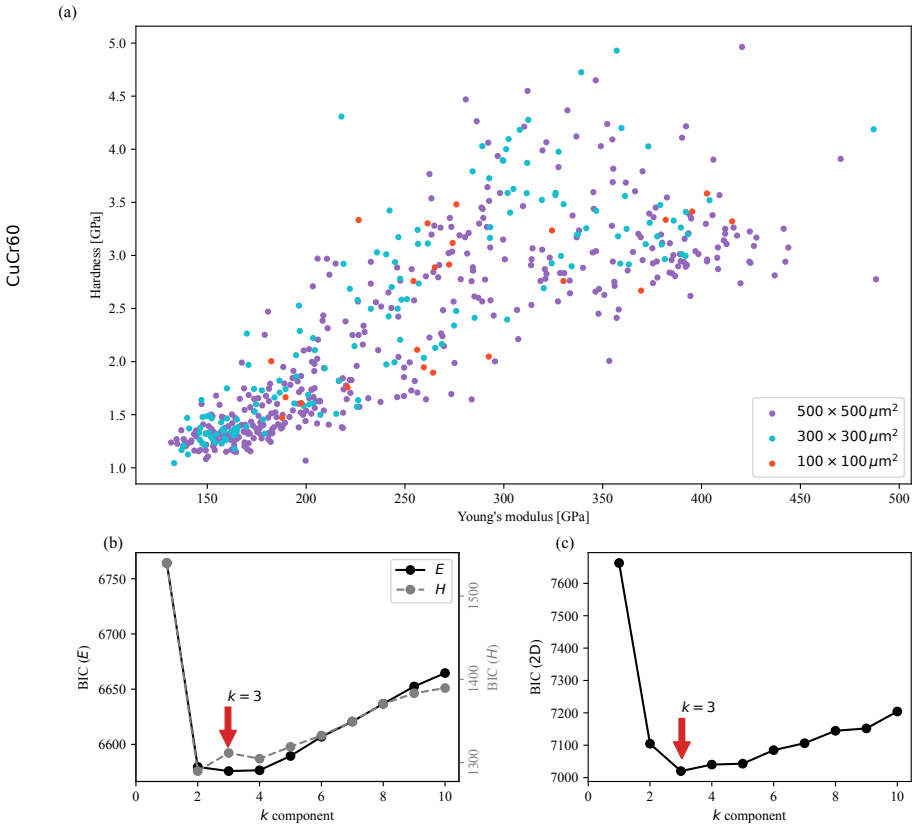

Figure 7.7. Depth-dependent distribution and BIC analysis of the CuCr60 data. (a) Distribution of the CuCr60 data over the depth range of 500 nm ~ 2000 nm (combined data for indentation areas of 100 × 100 µm², 300 × 300 µm² and 500 × 500 µm², respectively). (b) 1D BIC for E and H. (c) 2D BIC for E and H.

As shown in Figure 7.7b, the most likely number of phases determined by analyzing Young's Modulus is three, while two phases are most probable when analyzing the hardness. The 2D fit, though, also indicates $k = 3$, as can be seen in Figure 7.7c. The individual 2D fits for the different areas of 300 × 300 µm² and 500 × 500 µm², each reflect three mechanical phases with 67.1 vol% Cr and 59.5 vol% Cr, respectively. Not

Table 7.3. 1D mechanical property fitting of CuCr60 based on the optimal BIC at 500 nm ~ 2000 nm depth for three indentation arrays

Indentation area (μm²)	No.	Mechanical property	Average (GPa)	Standard deviation (GPa)	Percentage (%)
100 × 100	23	E_9		—	
		H_1	1.83	0.21	39.1
		H_2	3.15	0.29	60.9
300 × 300	154	E_1	157.79	12.02	29.2
		E_2	280.30	70.47	70.8
		H_1	1.42	0.18	37.3
		H_2	3.02	0.74	62.7
500 × 500	394	E_1	177.86	24.92	49.5
		E_2	275.26	31.88	24.9
		E_3	378.94	35.93	24.6
		H_1	1.44	0.22	45.5
		H_2	2.95	0.62	54.5
merged data	571	E_1	175.12	24.10	46.0
		E_2	269.55	34.10	28.5
		E_3	373.37	37.46	25.5
		H_1	1.45	0.22	42.8
		H_2	2.98	0.65	57.2

Table 7.4. 2D mechanical property fitting of CuCr60 based on the optimal BIC at 500 nm ~ 2000 nm depth for three indentation arrays

Indentation area (μm²)	Average E (GPa)	Average H (GPa)	Pct. (%)
100 × 100	252.19	2.46	78.4
	393.06	3.27	21.6
300 × 300	159.47	1.39	32.8
	229.10	2.29	31.9
	337.91	3.51	35.2
500 × 500	172.25	1.41	40.5
	267.99	2.74	38.1
	383.12	3.01	21.4
merged data	170.49	1.42	38.5
	264.73	2.78	40.4
	379.28	3.70	21.1

unexpectedly, the merged data then yielded a Cr fraction of 61.5 vol%. The somewhat different results reflect the variation of the microstructure over the sample surface and underscores the importance of identifying the characteristic microstructures or increasing the size of the dataset. How should one decide if this is a two or three phase system? From the point of materials science, one would probably choose only two phases, however, the analysis showed, that the properties of the third "phase" are rather distinct and can be recognized as a separate one.

7.3.4. Cross-Validation of Gaussian Mixture Model Results

The amount of required training data has been an often encountered concern in clustering algorithms for machine learning. In contrast to the model selection criteria used in the GMM with BIC, we will now focus on the robustness and validity of the clustering results as a function of the size of the dataset. Clustering together with cross validation is used to evaluate the effect of data size on the results and to identify the amount of experimental data required to achieve the same level of performance. We applied the following procedure:

1. Given the whole dataset $\mathcal{D}$, the categorical variable y_i is generated by a clustering algorithm A_k. It constructs such a solution $\boldsymbol{Y} \coloneqq A_k(\mathcal{D})$, where $\boldsymbol{Y} = \{y_1, ..., y_i\}$, and $y_i \in \{1, \ldots, k\}$ is assigned to the cluster. In our case, labels y_i were generated using the K-means and GMM clustering algorithms, with the prediction of the optimal GMM algorithm serving as the ground truth.

2. k-fold cross-validation is conducted on the dataset $\mathcal{D}$: First, divide the dataset into k equal parts (the "folds"), then choose $(k-1)$ folds for training and the remaining fold for testing. Then, validate the model using the testing dataset $\mathcal{D}_{\text{test}}$ after training the model with the training dataset $\mathcal{D}_{\text{train}}$. Perform k rounds of cross-validation using multiple training datasets. Note, that k indicates the number of folds and not the number of clusters as above.

3. The adjusted Rand index (ARI) is used to represent the performance level of the clustering algorithms. In general, the ARI approaches 0.0 for random clustering, reaches 1.0 for identical clusterings, but can go as low as -0.5 for highly discordant clusterings [182].

Figure 7.8a illustrates the outcomes of k-fold cross-validation conducted on 364 sets of CuCr60 data ($500 \times 500\,\mu m^2$) using two distinct clustering techniques, i.e., K-means and GMM. The GMM algorithm consistently outperformed K-means, as evidenced by the results. Figure 7.8b presents the results of k-fold cross-validation using the GMM algorithm across varying data sizes for CuCr60.

For the CuCr60 composite, a total of 571 datapoints were collected by merging data from three different indentation areas. We developed an approach to assess varying sizes of datasets, starting with taking 50 data records as the dataset of interest, and increasing by increments of 50 until reaching a dataset containing 550 data points.

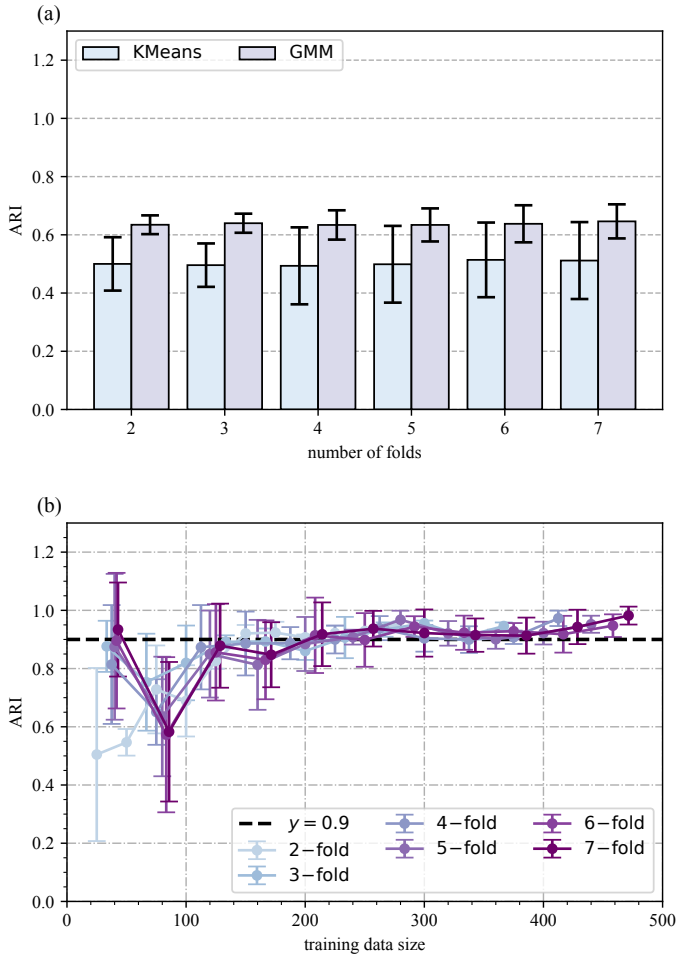

Figure 7.8. k-fold cross validation results of CuCr60. (a) The performance of two unsupervised models, K-means and GMM, measured by the ARI value and as a function of the number of folds of the k-fold cross-validation. (b) ARI values of k-fold cross-validation using GMM with varying the amount of training data.

During the process of training a model, a fraction of $(k-1)/k$ of the entire dataset was used as training data $\mathcal{D}_{\text{train}}$. For example, for 5-fold cross validation ($k = 5$), this entailed using 40, 80, 120, ... 440 datapoints for training, accompanied by 10, 20, 30, ... 60 datapoints for testing the model. During k-fold cross-validation, unique selections of training data were made, resulting in a total of five iterations for each fixed amount of datapoints. This iterative approach was adopted to gain insights into model performance concerning various data quantities, while minimizing any inaccuracies that may arise due to sampling bias.

Overall, our analysis indicates that the GMM algorithm exhibits superior accuracy when the average model performance approaches 0.9. Nevertheless, in the pursuit of enhanced generalizability, we aim to minimize errors further, ideally pushing the lower boundary of performance beyond 0.9. To achieve this, it is recommended to initiate the training process using a dataset comprising more than 400 datapoints as a starting point, as it appears to yield more reliable results.

The identical approach for sampling was employed in the examination of CuCr25. Figure 7.9 shows the outcomes of k-fold cross validation of CuCr25 with different numbers of clusters. Based on the 2D GMM analysis described above, the optimal model contained three clusters. However, we also notice that the BIC values determined for four and five components are comparable to those determined for three (Figure 7.6c). When the BIC values do not indicate a significant difference, the k-fold cross-validation can certainly provide a hint as to which model is superior. In the case of three clusters, the ARI values increased with the amount of data, whereas with other numbers of clusters the values did not exhibit an upward trend in conjunction with predicted values lower than 0.8. As a supplement to BIC, the k-fold cross-validation method can be used to determine how robust a certain number of clusters is and, more importantly, whether the amount of data is sufficient for training.

7.3.5. Effect of Indentation Depth

Figure 7.10a shows the 2D mapping of E for the CuCr25 dataset containing indents in the depth range between 200 nm and 600 nm. White color indicates positions for which no nanoindentation data was collected, blue represents the Cu-rich zones, and red represents the Cr-rich zones with color gradients in the vicinity of phase boundaries. Figure 7.10b shows the 1D Gaussian mixture model fitting results for Young's modulus indicating that the microstructure primarily consists of three mechanical phases, with the third phase mainly located at the intersection of the Cu-rich and Cr-rich regions. Figure 7.10c displays the 2D Gaussian mixture model's fitted results, with an additional fourth term.

The clustering results assign fractions of the Cu-phase at different depths to the fourth mechanical phase. Therefore, datasets containing data at different depths, need to be evaluated in subsets when determining the number of mechanical phases. Accordingly, we further examined the phase composition using a 1D Gaussian mixing model for different depth ranges covering 100 nm each. The results are shown in Table 7.5. Based on this analysis, the occurrence of additional phases can be observed

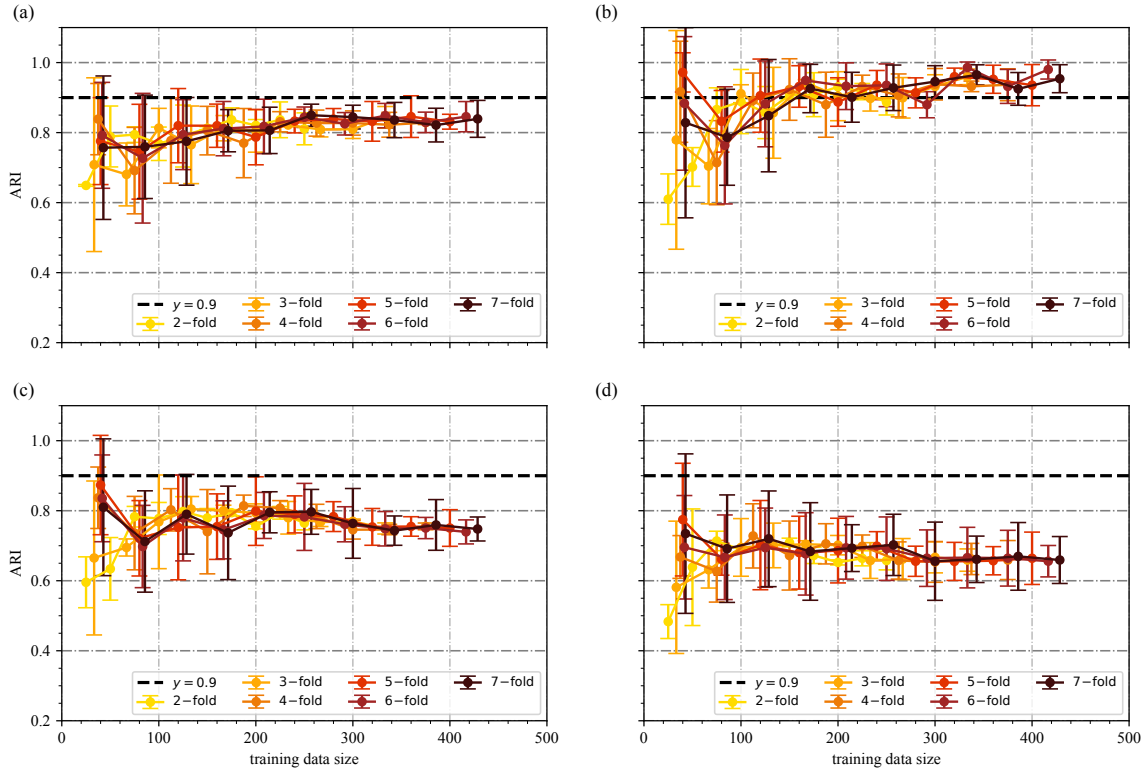

Figure 7.9. k-fold cross validation results of CuCr25 with different numbers of clusters for the GMM model and data size. (a) Two components. (b) Three components. (c) Four components. (d) Five components. A total of 513 datapoints were collected for the CuCr25 at 1 µm depth. The proportion of datapoints used as training data in a k-fold cross-validation is determined by the value of k.

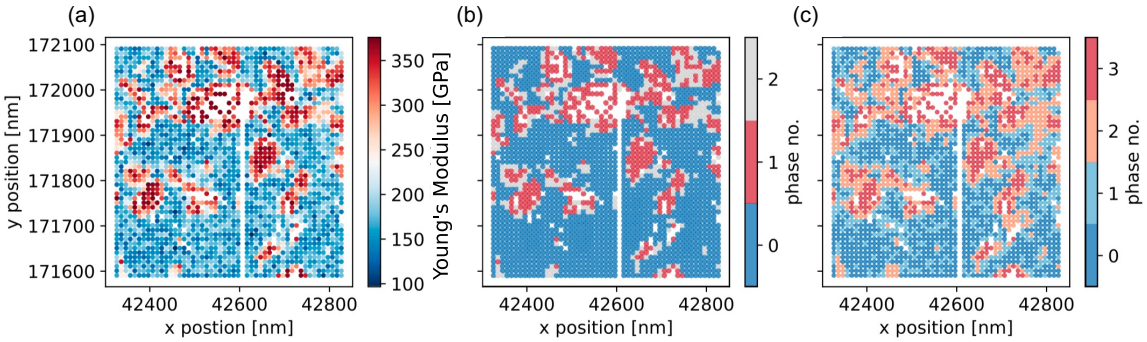

Figure 7.10. Elastic modulus of CuCr25 based on idents in the depth range of 200 nm ~ 600 nm. (a) Young's modulus distribution over the testing area. (b) Young's modulus distribution of 1D GMM clustering. (c) 2D GMM clustering results ($k = 4$).

at higher depths, i.e., in the 400 nm and 500 nm regimes, along with a significant reduction in Cu content.

The hardness distribution for the CuCr25 and CuCr60 composites at depths ranging from 200 nm to 1200 nm is shown in Figure 7.11. The data was grouped in 50 bins of equal size with the shaded area indicating the standard deviation. For metals typically an indentation size effect is observed, i.e., the hardness increases with decreasing indentation depth [183, 184, 78]. This indentation size effect is observed here for indentation depths between 450 nm and 600 nm, while for lower depths it is obscured by the combined effects of microstructural changes over the depth as well as the presence of at least two phases. This apparent nanoscale softening may also be influenced by the distance between indents relative to the size of the different phases possibly favoring a particular property value.

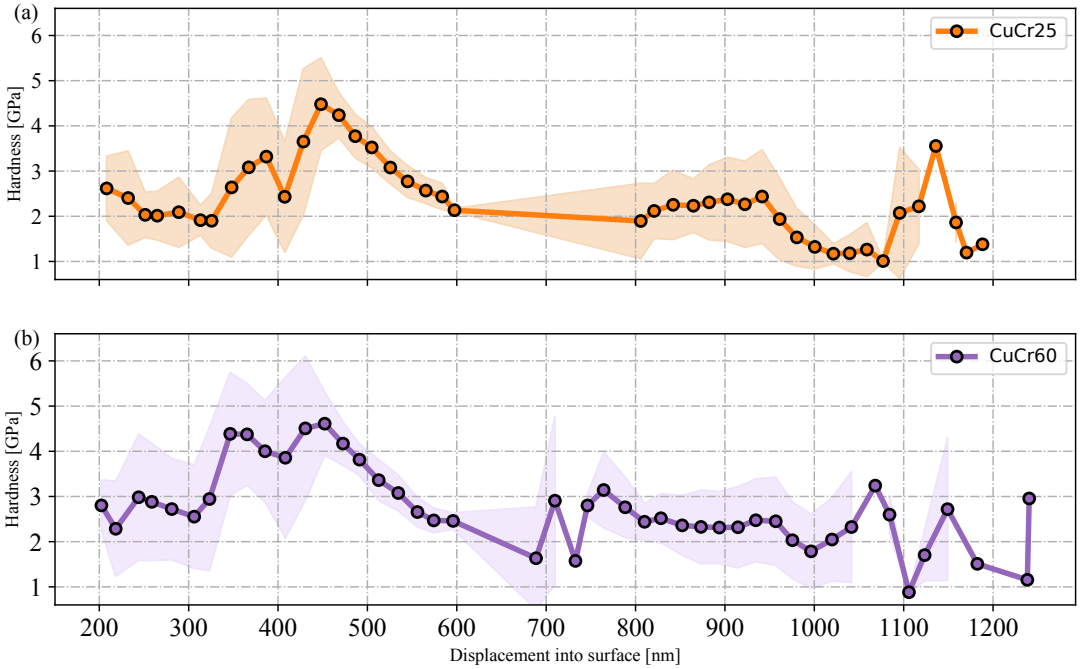

Figure 7.11. Hardness of CuCr25 and CuCr60 composites at different indentation depths. (a) CuCr25, with 3812 data points in the range 200 nm ~ 600 nm. (b) CuCr60, with 3254 data points in the range 200 nm ~ 600 nm depth. The mean of the hardness with the respective standard deviation as shaded area is shown. The mean hardness assumes an approximately constant value for depths greater than 600 nm.

Figure 7.12 depicts the distribution of Young's modulus in CuCr25 in the range between 250 nm ~ 600 nm, whose 2D top view is depicted in Figure 7.10a. The distribution of Young's modulus exhibits a certain degree of continuity, and the majority of the changes are gradual. The green hue represents the Cu-phase, while the orange color represents the Cr-phase. Observation reveals that the yellow and light-green portions (the third mechanical phase) can be considered the Cu and Cr intersection region.

Table 7.5. Depth-dependent Young's modulus fitting results for CuCr25 and CuCr60 based on the optimal BIC

Displacement into surface (nm)	CuCr25				CuCr60			
	Young's Modulus	Average (GPa)	Std Dev (GPa)	Pct. (%)	Young's Modulus	Average (GPa)	Std Dev (GPa)	Pct. (%)
200 ~ 300	E_1	163.72	23.81	95.2	E_1	162.10	27.37	76.2
	E_2	272.26	46.67	4.8	E_2	274.61	38.95	23.8
300 ~ 400	E_1	163.55	24.29	85.8	E_1	161.90	29.28	34.1
	E_2	290.88	38.88	14.2	E_2	296.06	36.07	65.9
400 ~ 500	E_1	158.35	19.65	13.4	E_1	182.94	36.95	14.7
	E_2	249.79	25.65	22.4	E_2	275.25	24.17	47.8
	E_3	303.13	18.81	32.9	E_3	326.85	21.85	37.5
500 ~ 600	E_1	351.83	15.34	31.3	E_1	198.21	31.98	69.7
	E_2	176.90	21.06	50.7	E_2	307.28	36.63	30.3
	E_3	242.53	30.22	25.2				
	E_4	341.30	24.67	24.1				

Figure 7.12 depicts the marker size adjusted according to the indent size on the surface, which increases as the indentation depth increases. Based on this observation and also on the analysis results listed in Table 7.5, we observe an increase of the Crphase in both the CuCr25 and the CuCr60 composites in the range of 200 nm ~ 500 nm, which can explain the increase in hardness in the low-depth regimes (Figure 7.11). Thus, it is critical to identify the characteristic microstructures not only on the surface tested but also in the sub-surface region of the material.

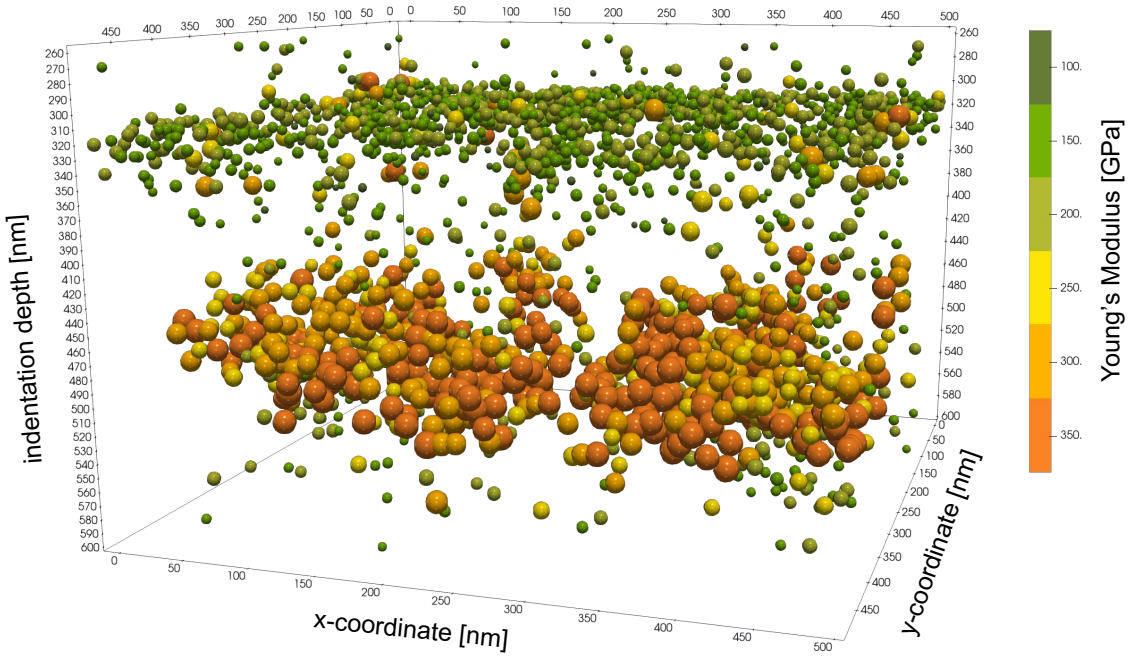

Figure 7.12. Spatial distribution of Young's modulus in CuCr25 as a function of indentation depth. The color of the spherical markers represents the Young's modulus.

7.4. Conclusion

Cu–Cr composites were studied by indentation as a model material to evaluate the ability to determine the properties of individual phases as well as the number of phases present. 1D and 2D Gaussian mixture models were trained and the most likely number of components identified based on the BIC. Using cross-validation, we demonstrated that GMM achieved more accurate results than K-means clustering. Investigating the dependence of the cross validation results on the size of the datasets helped to understand what a reasonable amount of data for the training of such models might be. Our analysis revealed that 450 data-pairs were sufficient for accurate phase volume prediction.

Clearly, the presented GMM-based method has limitations, e.g., it implicitly assumes that indents into different phases also result in different properties. As a consequence,

it must be expected that the predictions are getting worse the more similar two (or more) phases become. GMM is a simple yet robust method, which is not able to utilize all available microstructural information. For example, locality effects due to the fact that nearby indents may be correlated cannot be captured. Extending GMM, e.g., in terms of engineered features, could be a possible solution.

While large datasets produced in a very short time are quite impressive, also conventional nanoindentation approaches with typically fewer data can be used. Variations of the phase composition over the depth as well as the distance between indents relative to the size of the microstructural features may favor a particular phase. To avoid such bias indentation depth and characteristic microstructures need to be critically evaluated.

Part IV.

SUMMARY AND OUTLOOK

8. Summary and Outlook

8.1. Summary and Contributions

This thesis has developed a data-driven framework for rapid material characterization of metallic materials, integrating experimental observation, computational modeling, and advanced data analysis into a unified workflow. The motivation for this work stems from the increasing complexity and volume of materials data and the limitations of conventional approaches in resolving microstructural phenomena and linking them to macroscopic properties. By embracing the data-driven paradigm, the studies conducted here demonstrate how quantitative insights can be extracted from complex datasets, how structure–property relationships can be revealed more effectively, and how the overall process of materials characterization can be accelerated.

The central objective of this work was to overcome key challenges associated with microstructural characterization. These include the difficulty of capturing complex spatiotemporal features beyond the reach of conventional experimental techniques, the challenge of integrating heterogeneous data into interpretable models, and the question of how much data unsupervised machine learning models require to achieve robust clustering results. Addressing these issues required an interdisciplinary strategy that integrates high-resolution experiments, computational modeling, and probabilistic analysis into a coherent framework. This thesis demonstrates that such an approach can yield a deeper understanding of deformation processes and material behavior while enabling more efficient data utilization.

One of the core contributions of this work was the development of a methodology to use dislocations themselves as quantitative probes of the local energy landscape. Dislocations carry essential information about the microstructural environment they traverse, and their motion reflects interactions with obstacles that govern plastic deformation. By applying spatiotemporal data-mining techniques to *in situ* transmission electron microscopy observations, this study was able to identify, quantify, and statistically characterize pinning points that impede dislocation glide in a Cantor alloy. The analysis revealed several key insights: the strength of pinning points evolves as dislocations pass, some becoming weaker while others strengthen; their spatial position can shift in the direction of the Burgers vector; and their distribution is heterogeneous, consisting of both isolated and closely packed obstacles. Furthermore, to ensure methodological rigor, the 3D reconstruction workflow was benchmarked against several state-of-the-art reconstruction approaches, confirming both its accuracy and robustness. These results provide a new perspective on the microscopic origins of hardening and offer a robust framework for studying dislocation–obstacle interactions

in complex alloys. More broadly, the approach demonstrates how data-driven coarse-graining can transform qualitative microscopy observations into quantitative insights about the underlying physics.

The second major contribution focused on the integration of machine learning into microstructure-property analysis and, in particular, on determining the amount of data necessary for robust probabilistic modeling. Using nanoindentation experiments on Cu–Cr composites, a Gaussian mixture model was employed to analyze large datasets containing hundreds of measurements of hardness and modulus. This approach enabled the identification of distinct mechanical phases and their respective volume fractions directly from experimental data. Furthermore, a cross-validation strategy was introduced to assess data sufficiency, revealing that approximately 450 data pairs were adequate for reliable determination of phase volume fractions. This result is significant because it provides a practical guideline for experimental design, demonstrating that predictive accuracy can be achieved without excessive data collection. The study illustrates how probabilistic modeling and analysis of data sufficiency can support the broader adoption of machine learning techniques in materials science, especially in scenarios where data acquisition is time-consuming or costly.

Taken together, the studies presented in this thesis show that data-driven approaches can significantly improve the speed and precision of materials characterization, enabling its application from the nanoscale to the macroscopic level. These improvements are achieved by transforming dislocation motion into a quantitative probe of the microstructural energy landscape and by providing a systematic framework for evaluating data sufficiency in machine learning, thereby contributing both methodological innovations and practical insights to the field. Moreover, the framework developed here is general and adaptable, and its application is not limited to the specific materials or experimental conditions studied. It provides a foundation for future research into a wide range of phenomena, from defect interactions and phase transformations to the optimization of mechanical performance in metallic materials.

8.2. Outlook

Despite these advances, several challenges and limitations remain. The quality and representativeness of the input data still play a critical role in the performance of data-driven methods, and issues such as measurement noise, spatial resolution limits, and sampling bias can affect the reliability of the results. The Gaussian mixture model, while robust and interpretable, assumes that distinct mechanical phases exhibit sufficiently different properties and does not account for spatial correlations between nearby measurements. Similarly, the coarse-grained analysis of dislocation dynamics captures statistical trends but does not resolve atomic-scale mechanisms or capture the full complexity of chemical interactions at pinning sites. Addressing these limitations will require the integration of richer datasets, more sophisticated machine learning architectures, and deeper coupling between data-driven and physics-based models.

Looking ahead, there are several promising avenues for future work. One direction

is the integration of multimodal data, combining microscopy, spectroscopy, diffraction, and mechanical testing, to construct a more comprehensive picture of microstructure–property relationships. Another direction is the development of hybrid approaches that embed physical constraints into data-driven models, thereby enhancing their reliability, interpretability, and consistency with fundamental physical principles while reducing their reliance on large datasets. Extending spatiotemporal analysis to larger datasets and more complex loading conditions would provide further insight into dynamic strengthening mechanisms and their role in deformation behavior. Additionally, coupling the presented approaches with discrete dislocation dynamics or crystal plasticity simulations could enable powerful cross-validation and multiscale modeling strategies. Finally, standardizing data formats, establishing open repositories, and promoting FAIR (Findable, Accessible, Interoperable, Reusable) data practices will be crucial to accelerating the adoption and impact of data-centric approaches in materials science.

In conclusion, this thesis demonstrates that data-driven methods have the potential to fundamentally transform how material characterization is conducted in metallic materials. By bridging experimental observation with data-driven analysis and computational modeling, these approaches allow for the rapid extraction of meaningful patterns, the efficient utilization of data, and the discovery of structure–property relationships that were previously inaccessible. The results underscore the growing importance of data-driven strategies in materials research and point toward a future in which materials discovery and performance optimization are enabled by the iterative integration of data-driven characterization, physics-based modeling and simulation, and targeted experimental investigation.

Bibliography

[1] N. R. Council. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. The National Academies Press, 2008. ISBN: 978-0309119993. DOI: https://doi.org/10.17226/12199.

[2] R. W. Cahn and P. Haasen. *Physical metallurgy*. Vol. 2. Elsevier, 1996. ISBN: 978-0444898753.

[3] A. Agrawal and A. Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. In: *Apl Materials* 45:053208 (2016), 053208. DOI: https://doi.org/10.1063/1.4946894.

[4] S. Sandfeld and G. Po. Microstructural comparison of the kinematics of discrete and continuum dislocations models. In: *Modelling and Simulation in Materials Science and Engineering* 238:085003 (2015), 085003. DOI: https://doi.org/10.1088/0965-0393/23/8/085003.

[5] T. S. Kuhn and I. Hacking. *The structure of scientific revolutions*. Vol. 2. 2. University of Chicago press Chicago, 1970. ISBN: 978-0226458083.

[6] Maplesoft. *Galileo's Inclined Plane Experiment*. Accessed: 2025-09-22. 2018. URL https://www.maplesoft.com/support/help/maple/view.aspx?path=MathApps%2FGalileosInclinedPlaneExperiment.

[7] J. Callister William D. and D. G. Rethwisch. *Materials science and engineering: an introduction*. John wiley & sons, 2020. ISBN: 978-1119405498.

[8] D. B. Williams and C. B. Carter. The transmission electron microscope. In: *Transmission electron microscopy: a textbook for materials science*. Springer, 1996, pp. 3–17. ISBN: 978-0306453243. DOI: https://doi.org/10.1007/978-1-4757-2519-3_1.

[9] A. C. Fischer-Cripps. *Nanoindentation*. Springer, 2011. ISBN: 978-1441998712. DOI: https://doi.org/10.1007/978-1-4419-9872-9.

[10] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics*. Springer, 1963. ISBN: 978-0465023820.

[11] S. J. Blundell and K. M. Blundell. *Concepts in thermal physics*. Oxford University Press, 2010. ISBN: 978-0199562107.

[12] D. Hull and D. J. Bacon. *Introduction to dislocations*. Liverpool: Elsevier, 2011. ISBN: ISBN: 978-0080966724. DOI: https://doi.org/10.1016/s1369-7021(11)70217-6.

[13] P. M. Anderson, J. P. Hirth, and J. Lothe. *Theory of dislocations*. Cambridge University Press, 2017. ISBN: 978-0521864367.

[14] P. J. Denning, D. E. Comer, D. Gries, M. C. Mulder, A. Tucker, A. J. Turner, and P. R. Young. Computing as a discipline. In: *Computer* 222:63–70 (1989), 63–70. DOI: https://doi.org/10.1109/2.19833.

[15] R. LeSar. *Introduction to computational materials science: fundamentals to applications*. Cambridge University Press, 2013. DOI: https://doi.org/10.1017/CBO9781139033398.

[16] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. In: *Physical review* 1404A:A1133 (1965), A1133. DOI: https://doi.org/10.1103/PhysRev.140.A1133.

[17] D. Frenkel and B. Smit. *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2023. ISBN: 978-0323902922.

[18] O. C. Zienkiewicz and R. L. Taylor. *The Finite Element Method Set*. Elsevier, 2005. ISBN: 978-0750664318.

[19] T. Hey, S. Tansley, K. M. Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*. Vol. 1. Microsoft research Redmond, WA, 2009. ISBN: 978-0982544204. URL https://www.microsoft.com/en-us/research/wp-content/uploads/2009/10/Fourth_Paradigm.pdf.

[20] Y. Liu, C. Niu, Z. Wang, Y. Gan, Y. Zhu, S. Sun, and T. Shen. Machine learning in materials genome initiative: A review. In: *Journal of Materials Science & Technology* 57:113–122 (2020), 113–122. DOI: https://doi.org/10.1016/j.jmst.2020.01.067.

[21] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. Springer, 2006. ISBN: 978-0387310732.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. ISBN: 978-0387848570. DOI: https://doi.org/10.1007/978-0-387-84858-7.

[23] National Science and Technology Council (US). *Materials Genome Initiative for Global Competitiveness*. Executive Office of the President, National Science and Technology Council, 2011. URL https://www.mgi.gov/sites/mgi/files/materials_genome_initiative-final.pdf.

[24] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. In: *APL materials* 11 (2013). DOI: https://doi.org/10.1063/1.4812323.

[25] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, et al. AFLOW: An automatic framework for high-throughput materials discovery. In: *Computational Materials Science* 58:218–226 (2012), 218–226. DOI: https://doi.org/10.1016/j.commatsci.2012.02.005.

[26] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. In: *npj Computational Materials* 11:1–15 (2015), 1–15. DOI: https://doi.org/10.1038/npjcompumats.2015.10.

[27] C. Draxl and M. Scheffler. The NOMAD laboratory: from data sharing to artificial intelligence. In: *Journal of Physics: Materials* 23:036001 (2019), 036001. DOI: https://doi.org/10.1088/2515-7639/ab13bb.

[28] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann. Industry 4.0. In: *Business & information systems engineering* 64:239–242 (2014), 239–242. DOI: https://doi.org/10.1007/s12599-014-0334-4.

[29] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh. Machine learning for molecular and materials science. In: *Nature* 5597715:547–555 (2018), 547–555. DOI: https://doi.org/10.1038/s41586-018-0337-2.

[30] J. Schmidt, M. R. Marques, S. Botti, and M. A. Marques. Recent advances and applications of machine learning in solid-state materials science. In: *npj computational materials* 51:83 (2019), 83. DOI: https://doi.org/10.1038/s41524-019-0221-0.

[31] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim. Machine learning in materials informatics: recent applications and prospects. In: *npj Computational Materials* 31:54 (2017), 54. DOI: https://doi.org/10.1038/s41524-017-0056-5.

[32] D. Morgan and R. Jacobs. Opportunities and challenges for machine learning in materials science. In: *Annual Review of Materials Research* 50:71–103 (2020), 71–103. DOI: https://doi.org/10.1146/annurev-matsci-070218-010015.

[33] J. Frenkel. Zur theorie der elastizitätsgrenze und der festigkeit kristallinischer körper. In: *Zeitschrift für Physik* 377-8:572–609 (1926), 572–609. DOI: https://doi.org/10.1007/bf01397292.

[34] G. I. Taylor. The mechanism of plastic deformation of crystals. Part I.—Theoretical. In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 145855:362–387 (1934), 362–387. DOI: https://doi.org/10.1098/rspa.1934.0106.

[35] E. Orowan. Zur kristallplastizität. III: Über den mechanismus des gleitvorganges. In: *Zeitschrift für Physik* 899:634–659 (1934), 634–659. DOI: https://doi.org/10.1007/BF01341480.

[36] M. Polanyi. Lattice distortion which originates plastic flow. In: *Z. Phys* 899-10:660–662 (1934), 660–662.

[37] J. Burgers. Some considerations on the fields of stress connected with dislocations in a regular crystal lattice. In: *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*. Vol. 42. 1939, pp. 293–378.

[38] J. M. Burgers. Geometrical considerations concerning the structural irregularities to be assumed in a crystal. In: *Proceedings of the Physical Society* 521:23 (1940), 23. DOI: https://doi.org/10.1088/0959-5309/52/1/304.

[39] M. F. Ashby, H. Shercliff, and D. Cebon. *Materials: engineering, science, processing and design*. Butterworth-Heinemann, 2018. ISBN: 978-0081023761. DOI: https://doi.org/10.1016/C2016-0-04242-0.

[40] A. H. Cottrell and M. Jaswon. Distribution of solute atoms round a slow dislocation. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 1991056:104–114 (1949), 104–114. DOI: https://doi.org/10.1098/rspa.1949.0128.

[41] J. Eshelby. Uniformly moving dislocations. In: *Proceedings of the Physical Society. Section A* 625:307 (1949), 307. DOI: https://doi.org/10.1088/0370-1298/62/5/307.

[42] W. Shockley and W. Read. Quantitative predictions from dislocation models of crystal grain boundaries. In: *Physical Review* 754:692 (1949), 692. DOI: https://doi.org/10.1103/physrev.75.692.

[43] W. T. Read and W. Shockley. Dislocation models of crystal grain boundaries. In: *Physical review* 783:275 (1950), 275. DOI: https://doi.org/10.1103/physrev.78.275.

[44] F. C. Frank and J. Read W. T. Multiplication processes for slow moving dislocations. In: *Physical Review* 794:722 (1950), 722. DOI: https://doi.org/10.1103/physrev.79.722.

[45] F. Vogel, W. Pfann, H. Corey, and E. Thomas. Observations of dislocations in lineage boundaries in germanium. In: *Physical Review* 903:489 (1953), 489. DOI: https://doi.org/10.1103/physrev.90.489.

[46] P. Hirsch, R. Horne, and M. Whelan. LXVIII. Direct observations of the arrangement and motion of dislocations in aluminium. In: *Philosophical Magazine* 17:677–684 (1956), 677–684. DOI: https://doi.org/10.1080/14786435608244003.

[47] E. O. Hall. The deformation and ageing of mild steel: III discussion of results. In: *Proceedings of the Physical Society. Section B* 649:747 (1951), 747. DOI: https://doi.org/10.1088/0370-1301/64/9/303.

[48] N. J. Petch. The cleavage strength of polycrystals. In: *J. Iron Steel Inst.* 174:25–28 (1953), 25–28.

[49] V. Volterra. Sur l'équilibre des corps élastiques multiplement connexes. In: *Annales scientifiques de l'École normale supérieure*. Vol. 24. 1907, pp. 401–517. DOI: https://doi.org/10.24033/asens.583.

[50] J. P. Hirth, J. Lothe, and T. Mura. Theory of dislocations. In: *Journal of Applied Mechanics* 502:476–477 (1983), 476–477. DOI: https://doi.org/10.1115/1.3167075.

[51] J. W. Yeh, S. K. Chen, S. J. Lin, J. Y. Gan, T. S. Chin, T. T. Shun, C. H. Tsau, and S. Y. Chang. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. In: *Advanced Engineering Materials* 65:299–303 (2004), 299–303. DOI: https://doi.org/10.1002/adem.200300567.

[52] B. Cantor, I. Chang, P. Knight, and A. Vincent. Microstructural development in equiatomic multicomponent alloys. In: *Materials Science and Engineering: A* 375:213–218 (2004), 213–218. DOI: https://doi.org/10.1016/j.msea.2003.10.257.

[53] S. Gorsse, J.-P. Couzinié, and D. B. Miracle. From high-entropy alloys to complex concentrated alloys. In: *Comptes Rendus Physique* 198:721–736 (2018), 721–736. DOI: https://doi.org/10.1016/j.crhy.2018.09.004.

[54] Z. Wu, H. Bei, G. M. Pharr, and E. P. George. Temperature dependence of the mechanical properties of equiatomic solid solution alloys with face-centered cubic crystal structures. In: *Acta Materialia* 81:428–441 (2014), 428–441. DOI: https://doi.org/10.1016/j.actamat.2014.08.026.

[55] O. N. Senkov, D. B. Miracle, K. J. Chaput, and J.-P. Couzinie. Development and exploration of refractory high entropy alloys—A review. In: *Journal of materials research* 3319:3092–3128 (2018), 3092–3128. DOI: https://doi.org/10.1557/jmr.2018.153.

[56] M. Zhang, E. George, and J. Gibeling. Elevated-temperature Deformation Mechanisms in a CrMnFeCoNi High-Entropy Alloy. In: *Acta Materialia* 218:117181 (2021), 117181. DOI: https://doi.org/10.1016/j.actamat.2021.117181.

[57] P. Moretti, M.-C. Miguel, M. Zaiser, and S. Zapperi. Depinning transition of dislocation assemblies: Pileups and low-angle grain boundaries. In: *Physical Review B* 6921:214103 (2004), 214103. DOI: https://doi.org/10.1103/physrevb.69.214103.

[58] G. Laplanche, A. Kostka, O. Horst, G. Eggeler, and E. George. Microstructure evolution and critical stress for twinning in the CrMnFeCoNi high-entropy alloy. In: *Acta Materialia* 118:152–163 (2016), 152–163. DOI: https://doi.org/10.1016/j.actamat.2016.07.038.

[59] F. Otto, A. Dlouhỳ, C. Somsen, H. Bei, G. Eggeler, and E. P. George. The influences of temperature and microstructure on the tensile properties of a CoCrFeMnNi high-entropy alloy. In: *Acta Materialia* 6115:5743–5755 (2013), 5743–5755. DOI: https://doi.org/10.1016/j.actamat.2013.06.018.

[60] Z. Wu, Y. Gao, and H. Bei. Thermal activation mechanisms and Labusch-type strengthening analysis for a family of high-entropy and equiatomic solid-solution alloys. In: *Acta Materialia* 120:108–119 (2016), 108–119. DOI: https://doi.org/10.1016/j.actamat.2016.08.047.

[61] R. Labusch. A statistical theory of solid solution hardening. In: *physica status solidi (b)* 412:659–669 (1970), 659–669. DOI: https://doi.org/10.1002/pssb.19700410221.

[62] R. Zhang, S. Zhao, J. Ding, Y. Chong, T. Jia, C. Ophus, M. Asta, R. O. Ritchie, and A. M. Minor. Short-range order and its impact on the CrCoNi medium-entropy alloy. In: *Nature* 5817808:283–287 (2020), 283–287. DOI: https://doi.org/10.1038/s41586-020-2275-z.

[63] Z. Shen, J. P. Du, S. Shinzato, Y. Sato, P. Yu, and S. Ogata. Kinetic Monte Carlo simulation framework for chemical short-range order formation kinetics in a multi-principal-element alloy. In: *Computational Materials Science* 198:110670 (2021), 110670. DOI: https://doi.org/10.1016/j.commatsci.2021.110670.

[64] Q. He, P. Tang, H. Chen, S. Lan, J. Wang, J. Luan, M. Du, Y. Liu, C. Liu, C. Pao, et al. Understanding chemical short-range ordering/demixing coupled with lattice distortion in solid solution high entropy alloys. In: *Acta Materialia* 216:117140 (2021), 117140. DOI: https://doi.org/10.1016/j.actamat.2021.117140.

[65] Y. Wu, F. Zhang, X. Yuan, H. Huang, X. Wen, Y. Wang, M. Zhang, H. Wu, X. Liu, H. Wang, et al. Short-range ordering and its effects on mechanical properties of high-entropy alloys. In: *Journal of Materials Science and Technology* 62:214–220 (2021), 214–220. DOI: https://doi.org/10.1016/j.jmst.2020.06.018.

[66] S. Lee, M. J. Duarte, M. Feuerbacher, R. Soler, C. Kirchlechner, C. H. Liebscher, S. H. Oh, and G. Dehm. Dislocation plasticity in FeCoCrMnNi high-entropy alloy: quantitative insights from in situ transmission electron microscopy deformation. In: *Materials Research Letters* 86:216–224 (2020), 216–224. DOI: https://doi.org/10.1080/21663831.2020.1741469.

[67] S. Chen, H. S. Oh, B. Gludovatz, S. J. Kim, E. S. Park, Z. Zhang, R. O. Ritchie, and Q. Yu. Real-time observations of TRIP-induced ultrahigh strain hardening in a dual-phase CrMnFeCoNi high-entropy alloy. In: *Nature communications* 111:1–8 (2020), 1–8. DOI: https://doi.org/10.1038/s41467-020-14641-1.

[68] P. Wang, Y. Bu, J. Liu, Q. Li, H. Wang, and W. Yang. Atomic deformation mechanism and interface toughening in metastable high entropy alloy. In: *Materials Today* 37:64–73 (2020), 64–73. DOI: https://doi.org/10.1016/j.mattod.2020.02.017.

[69] C. Varvenne, A. Luque, and W. A. Curtin. Theory of strengthening in fcc high entropy alloys. In: *Acta Materialia* 118:164–176 (2016), 164–176. DOI: https://doi.org/10.1016/j.actamat.2016.07.040.

[70] Q. J. Li, H. Sheng, and E. Ma. Strengthening in multi-principal element alloys with local-chemical-order roughened dislocation pathways. In: *Nature communications* 101:1–11 (2019), 1–11. DOI: https://doi.org/10.1038/s41467-019-11464-7.

[71] Y. Bu, Y. Wu, Z. Lei, X. Yuan, H. Wu, X. Feng, J. Liu, J. Ding, Y. Lu, H. Wang, et al. Local chemical fluctuation mediated ductility in body-centered-cubic high-entropy alloys. In: *Materials Today* 46:28–34 (2021), 28–34. DOI: https://doi.org/10.1016/j.mattod.2021.02.022.

[72] M. Komarasamy, N. Kumar, R. S. Mishra, and P. K. Liaw. Anomalies in the deformation mechanism and kinetics of coarse-grained high entropy alloy. In: *Materials Science and Engineering: A* 654:256–263 (2016), 256–263. DOI: https://doi.org/10.1016/j.msea.2015.12.063.

[73] Y. Shen. Nanoindentation for Testing Material Properties. In: *Handbook of Mechanics of Materials*. Ed. by S. Schmauder, C. Chen, K. Chawla, N. Chawla, W. Chen, and Y. Kagawa. Singapore: Springer, 2019, pp. 1981–2012. ISBN: 978-9811068836. DOI: https://doi.org/10.1007/978-981-10-6884-3_46.

[74] Y. Golovin. Nanoindentation and Mechanical Properties of Materials at Submicro- and Nanoscale Levels: Recent Results and Achievements. In: *Phys. Solid State* 63:1–41 (2021), 1–41. DOI: https://doi.org/10.1134/s1063783421010108.

[75] H. Hertz. The contact of elastic solids. In: *J Reine Angew, Math* 92:156–171 (1881), 156–171.

[76] J. Brinell. Way of determining the hardness of bodies and some applications of the same. In: *Teknisk Tidskrift* 5:69 (1900), 69.

[77] R. L. Smith and G. Sandly. An accurate method of determining the hardness of metals, with particular reference to those of a high degree of hardness. In: *Proceedings of the Institution of Mechanical Engineers* 1021:623–641 (1922), 623–641. DOI: https://doi.org/10.1243/PIME_PROC_1922_102_033_02.

[78] G. M. Pharr, E. G. Herbert, and Y. Gao. The indentation size effect: a critical examination of experimental observations and mechanistic interpretations. In: *Annual Review of Materials Research* 40:271–292 (2010), 271–292. DOI: https://doi.org/10.1146/annurev-matsci-070909-104456.

[79] I. M. Hutchings. The contributions of David Tabor to the science of indentation hardness. In: *Journal of Materials Research* 243:581–589 (2009), 581–589. DOI: https://doi.org/10.1557/jmr.2009.0085.

[80] A. A. Griffith. VI. The phenomena of rupture and flow in solids. In: *Philosophical transactions of the royal society of london. Series A, containing papers of a mathematical or physical character* 221582-593:163–198 (1921), 163–198. DOI: https://doi.org/10.1098/rsta.1921.0006.

[81] D. Tabor. *The hardness of metals*. Oxford university press, 2000. ISBN: 0198507763.

[82] I. N. Sneddon. The relation between load and penetration in the axisymmetric Boussinesq problem for a punch of arbitrary profile. In: *International journal of engineering science* 31:47–57 (1965), 47–57. DOI: https://doi.org/10.1016/0020-7225(65)90019-4.

[83] W. C. Oliver and G. M. Pharr. An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. In: *Journal of materials research* 76:1564–1583 (1992), 1564–1583. DOI: https://doi.org/10.1557/JMR.1992.1564.

[84] W. C. Oliver and G. M. Pharr. Measurement of hardness and elastic modulus by instrumented indentation: Advances in understanding and refinements to methodology. In: *Journal of materials research* 191:3–20 (2004), 3–20. DOI: https://doi.org/10.1557/jmr.2004.0002.

[85] W. D. Nix and H. Gao. Indentation size effects in crystalline materials: a law for strain gradient plasticity. In: *Journal of the Mechanics and Physics of Solids* 463:411–425 (1998), 411–425. DOI: https://doi.org/10.1016/S0022-5096(97)00086-0.

[86] G. Constantinides, K. R. Chandran, F.-J. Ulm, and K. Van Vliet. Grid indentation analysis of composite microstructure and mechanics: Principles and validation. In: *Materials Science and Engineering: A* 4301-2:189–202 (2006), 189–202. DOI: https://doi.org/10.1016/j.msea.2006.05.125.

[87] J. Nohava, P. Haušild, Š. Houdková, and R. Enžl. Comparison of Isolated Indentation and Grid Indentation Methods for HVOF Sprayed Cermets. In: *J Therm Spray Tech* 21:651–658 (2012), 651–658. DOI: https://doi.org/10.1007/s11666-012-9733-6.

[88] C.-M. Sanchez-Camargo, A. Hor, M. Salem, and C. Mabru. A robust method for mechanical characterization of heterogeneous materials by nanoindentation grid analysis. In: *Materials & Design* 194:108908 (2020), 108908. DOI: https://doi.org/10.1016/j.matdes.2020.108908.

[89] H. Besharatloo and J. M. Wheeler. Influence of indentation size and spacing on statistical phase analysis via high-speed nanoindentation mapping of metal alloys. In: *Journal of Materials Research* 3611:2198–2212 (2021), 2198–2212. DOI: https://doi.org/10.1557/s43578-021-00214-5.

[90] J. Wheeler. Mechanical phase mapping of the Taza meteorite using correlated high-speed nanoindentation and EDX. In: *J. Mater. Res.* 361:94–104 (2021), 94–104. DOI: https://doi.org/10.1557/s43578-020-00056-7.

[91] B. Vignesh, W. Oliver, G. S. Kumar, and P. S. Phani. Critical assessment of high speed nanoindentation mapping technique and data deconvolution on thermal barrier coatings. In: *Materials & Design* 181:108084 (2019), 108084. DOI: https://doi.org/10.1016/j.matdes.2019.108084.

[92] C. Tromas, M. Arnoux, and X. Milhet. Hardness cartography to increase the nanoindentation resolution in heterogeneous materials: application to a Ni-based single-crystal superalloy. In: *Scr. Mater.* 662:77 (2012), 77. DOI: https://doi.org/10.1016/j.scriptamat.2011.09.042.

[93] E. Hintsala, U. Hangen, and D. Stauffer. High-Throughput Nanoindentation for Statistical and Spatial Property Determination. In: *JOM* 70:494–503 (2018), 494–503. DOI: https://doi.org/10.1007/s11837-018-2752-0.

[94] F.-J. Ulm, M. Vandamme, C. Bobko, J. A. Ortega, K. Tai, and C. Ortiz. Statistical indentation techniques for hydrated nanocomposites: concrete, bone, and shale. In: *J. Am. Ceram. Soc.* 909:2677 (2007), 2677. DOI: https://doi.org/10.1111/j.1551-2916.2007.02012.x.

[95] P. Haušild, A. Materna, L. Kocmanová, and J. Matějíček. Determination of the individual phase properties from the measured grid indentation data. In: *J. Mater. Res.* 3122:3538–3548 (2016), 3538–3548. DOI: https://doi.org/10.1557/jmr.2016.375.

[96] L. Karanja, M. Lenci, D. Piot, C. Maurice, A. Durif, M. Richou, L. Gallais, M. Minissale, and G. Kermouche. An attempt to assess recovery/recrystallization kinetics in tungsten at high temperature using statistical nanoindentation analysis. In: *Crystals* 111:37 (2020), 37. DOI: https://doi.org/10.3390/cryst11010037.

[97] E. S. Puchi-Cabrera, E. Rossi, G. Sansonetti, M. Sebastiani, and E. Bemporad. Machine learning aided nanoindentation: A review of the current state and future perspectives. In: *Current Opinion in Solid State and Materials Science* 274:101091 (2023), 101091. DOI: https://doi.org/10.1016/j.cossms.2023.101091.

[98] Y. B. Veytskin, V. K. Tammina, C. P. Bobko, P. G. Hartley, M. B. Clennell, D. N. Dewhurst, and R. R. Dagastine. Micromechanical characterization of shales through nanoindentation and energy dispersive X-ray spectrometry. In: *Geomech. Energy Environ.* 9:21–35 (2017), 21–35. DOI: https://doi.org/10.1016/j.gete.2016.10.004.

[99] L. Sorelli, G. Constantinides, F.-J. Ulm, and F. Toutlemonde. The nano-mechanical signature of ultra high performance concrete by statistical nanoindentation techniques. In: *Cement and Concrete Research* 3812:1447–1456 (2008), 1447–1456. DOI: https://doi.org/10.1016/j.cemconres.2008.09.002.

[100] N. X. Randall, M. Vandamme, and F.-J. Ulm. Nanoindentation analysis as a two-dimensional tool for mapping the mechanical properties of complex surfaces. In: *Journal of materials research* 243:679–690 (2009), 679–690. DOI: https://doi.org/10.1557/jmr.2009.0149.

[101] S. Kossman and M. Bigerelle. Pop-in identification in nanoindentation curves with deep learning algorithms. In: *Materials* 1422:7027 (2021), 7027. DOI: https://doi.org/10.3390/ma14227027.

[102] R. M. Jentner, K. Srivastava, S. Scholl, F.-J. Gallardo-Basile, J. P. Best, C. Kirchlechner, and G. Dehm. Unsupervised clustering of nanoindentation data for microstructural reconstruction: Challenges in phase discrimination. In: *Materialia* 28:101750 (2023), 101750. DOI: https://doi.org/10.1016/j.mtla.2023.101750.

[103] B. R. Becker, E. D. Hintsala, B. Stadnick, U. D. Hangen, and D. D. Stauffer. Automated analysis method for high throughput nanoindentation data with quantitative uncertainty. In: *Journal of Applied Physics* 13218:185101 (2022), 185101. DOI: https://doi.org/10.1063/5.0098493.

[104] S. R. Kalidindi and M. De Graef. Materials data science: current status and future outlook. In: *Annual Review of Materials Research* 45:171–193 (2015), 171–193. DOI: https://doi.org/10.1146/annurev-matsci-070214-020844.

[105] R. Bostanabad, Y. Zhang, X. Li, T. Kearney, L. C. Brinson, D. W. Apley, W. K. Liu, and W. Chen. Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. In: *Progress in Materials Science* 95:1–41 (2018), 1–41. DOI: https://doi.org/10.1016/j.pmatsci.2018.01.005.

[106] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific data* 31:1–9 (2016), 1–9. DOI: https://doi.org/10.1038/sdata.2016.18.

[107] M. Jalali, Y. Luo, L. Caulfield, E. Sauter, A. Nefedov, and C. Wöll. Large language models in electronic laboratory notebooks: Transforming materials science research workflows. In: *Materials Today Communications* 40:109801 (2024), 109801. DOI: https://doi.org/10.1016/j.mtcomm.2024.109801.

[108] R. P. Abernathey, T. Augspurger, A. Banihirwe, C. C. Blackmon-Luca, T. J. Crone, C. L. Gentemann, J. J. Hamman, N. Henderson, C. Lepore, T. A. McCaie, et al. Cloud-native repositories for big scientific data. In: *Computing in Science & Engineering* 232:26–35 (2021), 26–35. DOI: https://doi.org/10.1109/MCSE.2021.3059437.

[109] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke. Data-driven materials science: status, challenges, and perspectives. In: *Advanced Science* 621:1900808 (2019), 1900808. DOI: https://doi.org/10.1002/advs.201900808.

[110] J. O'Mara, B. Meredig, and K. Michel. Materials data infrastructure: a case study of the citrination platform to examine data import, storage, and access. In: *Jom* 688:2031–2034 (2016), 2031–2034. DOI: https://doi.org/10.1007/s11837-016-1984-0.

[111] D. S. Karls, M. Bierbaum, A. A. Alemi, R. S. Elliott, J. P. Sethna, and E. B. Tadmor. The OpenKIM processing pipeline: A cloud-based automatic material property computation engine. In: *The Journal of Chemical Physics* 1536 (2020). DOI: https://doi.org/10.1063/5.0014267.

[112] S. Sandfeld. *Materials Data Science: Introduction to Data Mining, Machine Learning, and Data-Driven Predictions for Materials Science and Engineering*. Springer Nature, 2024. ISBN: 978-3031465642. DOI: https://doi.org/10.1007/978-3-031-46565-9.

[113] S. R. Kalidindi. *Hierarchical materials informatics: novel analytics for materials data*. Elsevier, 2015. ISBN: 978-0124103948. DOI: https://doi.org/10.1016/C2012-0-07337-1.

[114] T. Bilyk, A. M. Goryaeva, M.-C. Marinica, C. Flament, C. Sabathier, E. Leroy, M. Loyer-Prost, and E. Meslin. Accurate quantification of dislocation loops in complex functional alloys enabled by deep learning image analysis. In: *Scientific Reports* 141:25168 (2024), 25168. DOI: https://doi.org/10.1038/s41598-024-74894-4.

[115] K. Govind, D. Oliveros, A. Dlouhy, M. Legros, and S. Sandfeld. Deep learning of crystalline defects from TEM images: A solution for the problem of 'never enough training data'. In: *Machine learning: science and technology* 51:015006 (2024), 015006. DOI: https://doi.org/10.1088/2632-2153/ad1a4e.

[116] B. L. DeCost, T. Francis, and E. A. Holm. Exploring the microstructure manifold: image texture representations applied to ultrahigh carbon steel microstructures. In: *Acta Materialia* 133:30–40 (2017), 30–40. DOI: https://doi.org/10.1016/j.actamat.2017.05.014.

[117] S. R. Spurgeon, C. Ophus, L. Jones, A. Petford-Long, S. V. Kalinin, M. J. Olszta, R. E. Dunin-Borkowski, N. Salmon, K. Hattar, W.-C. D. Yang, et al. Towards data-driven next-generation transmission electron microscopy. In: *Nature materials* 203:274–279 (2021), 274–279. DOI: https://doi.org/10.1038/s41563-020-00833-z.

[118] T. Hochrainer, S. Sandfeld, M. Zaiser, and P. Gumbsch. Continuum dislocation dynamics: towards a physical theory of crystal plasticity. In: *Journal of the Mechanics and Physics of Solids* 63:167–178 (2014), 167–178. DOI: https://doi.org/10.1016/j.jmps.2013.09.012.

[119] D. Steinberger, H. Song, and S. Sandfeld. Machine learning-based classification of dislocation microstructures. In: *Frontiers in Materials* 6:141 (2019), 141. DOI: https://doi.org/10.3389/fmats.2019.00141.

[120] H. Song, B. D. Nguyen, K. Govind, D. Berta, P. D. Ispanovity, M. Legros, and S. Sandfeld. Enabling quantitative analysis of in situ TEM experiments: A high-throughput, deep learning-based approach tailored to the dynamics of dislocations. In: *Acta Materialia* 282:120455 (2025), 120455. DOI: https://doi.org/10.1016/j.actamat.2024.120455.

[121] J. Wagner, C. G. Berger, X. Du, T. Stubhan, J. A. Hauch, and C. J. Brabec. The evolution of Materials Acceleration Platforms: toward the laboratory of the future with AMANDA. In: *Journal of Materials Science* 5629:16422–16446 (2021), 16422–16446. DOI: https://doi.org/10.1007/s10853-021-06281-7.

[122] M. Rout, S. K. Pal, and S. B. Singh. Cross rolling: a metal forming process. In: *Modern manufacturing engineering*:41–64 (2015), 41–64. DOI: https://doi.org/10.1007/978-3-319-20152-8_2.

[123] D. Rowenhorst, A. Rollett, G. Rohrer, M. Groeber, M. Jackson, P. J. Konijnenberg, and M. De Graef. Consistent representations of and conversions between 3D rotations. In: *Modelling and Simulation in Materials Science and Engineering* 238:083501 (2015), 083501. DOI: https://doi.org/10.1088/0965-0393/23/8/083501.

[124] H.-J. Bunge and W. Roberts. Orientation distribution, elastic and plastic anisotropy in stabilized steel sheet. In: *Applied Crystallography* 23:116–128 (1969), 116–128. DOI: https://doi.org/10.1107/s0021889869006704.

[125] F. Mompiou and R.-x. Xie. pycotem: An open source toolbox for online crystal defect characterization from TEM imaging and diffraction. In: *Journal of Microscopy* 2821:84–97 (2021), 84–97. DOI: https://doi.org/10.1111/jmi.12982.

[126] G. Team. *GeoGebra*. https://www.geogebra.org/. Computer software. 2025.

[127] J. Mayer, L. A. Giannuzzi, T. Kamino, and J. Michael. TEM sample preparation and FIB-induced damage. In: *MRS bulletin* 325:400–407 (2007), 400–407. DOI: https://doi.org/10.1557/mrs2007.63.

[128] S. Li, Y. Chang, Y. Wang, Q. Xu, and B. Ge. A review of sample thickness effects on high-resolution transmission electron microscopy imaging. In: *Micron* 130:102813 (2020), 102813. DOI: https://doi.org/10.1016/j.micron.2019.102813.

[129] F. Delmas, M.-J. Casanove, P. Lours, A. Couret, and A. Coujou. Quantitative TEM study of the precipitation microstructure in aluminium alloy Al (MgSiCu) 6056 T6. In: *Materials Science and Engineering: A* 3731-2:80–89 (2004), 80–89. DOI: https://doi.org/10.1016/j.msea.2003.12.068.

[130] D. Oliveros, A. Fraczkiewicz, A. Dlouhy, C. Zhang, H. Song, S. Sandfeld, and M. Legros. Orientation-related twinning and dislocation glide in a Cantor High Entropy Alloy at room and cryogenic temperature studied by in situ TEM straining. In: *Materials Chemistry and Physics* 272:124955 (2021), 124955. DOI: https://doi.org/10.1016/j.matchemphys.2021.124955.

[131] M. Peach and J. Koehler. The forces exerted on dislocations and the stress fields produced by them. In: *Physical Review* 803:436 (1950), 436. DOI: https://doi.org/10.1103/PhysRev.80.436.

[132] V. Bulatov and W. Cai. *Computer simulations of dislocations*. Vol. 3. Oxford University Press, 2006. ISBN: 9780198526148. DOI: https://doi.org/10.1093/oso/9780198526148.001.0001.

[133] W. Cai, A. Arsenlis, C. R. Weinberger, and V. V. Bulatov. A non-singular continuum theory of dislocations. In: *Journal of the Mechanics and Physics of Solids* 543:561–587 (2006), 561–587. DOI: https://doi.org/10.1016/j.jmps.2005.09.005.

[134] N. M. Ghoniem, S. H. Tong, and L. Z. Sun. Parametric dislocation dynamics: a thermodynamics-based approach to investigations of mesoscopic plastic deformation. In: *Physical Review B* 612:913 (2000), 913. DOI: https://doi.org/10.1103/physrevb.61.913.

[135] B. Devincre and L. P. Kubin. Mesoscopic simulations of dislocations and plasticity. In: *Materials Science and Engineering: A* 234:8–14 (1997), 8–14. DOI: https://doi.org/10.1016/s0921-5093(97)00146-9.

[136] D. Raabe. Introduction of a hybrid model for the discrete 3D simulation of dislocation dynamics. In: *Computational materials science* 111:1–15 (1998), 1–15. DOI: https://doi.org/10.1016/S0927-0256(97)00160-2.

[137] A. Arsenlis, W. Cai, M. Tang, M. Rhee, T. Oppelstrup, G. Hommes, T. G. Pierce, and V. V. Bulatov. Enabling strain hardening simulations with dislocation dynamics. In: *Modelling and Simulation in Materials Science and Engineering* 156:553 (2007), 553. DOI: https://doi.org/10.1088/0965-0393/15/6/001.

[138] C. Zhou, S. B. Biner, and R. LeSar. Discrete dislocation dynamics simulations of plasticity at small scales. In: *Acta Materialia* 585:1565–1577 (2010), 1565–1577. DOI: https://doi.org/10.1016/j.actamat.2009.11.001.

[139] V. Bulatov, W. Cai, J. Fier, M. Hiratani, G. Hommes, T. Pierce, M. Tang, M. Rhee, K. Yates, and T. Arsenlis. Scalable line dynamics in ParaDiS. In: *SC'04: Proceedings of the 2004 ACM/IEEE Conference on Supercomputing*. IEEE. 2004, pp. 19–19. DOI: https://doi.org/10.1109/SC.2004.53.

[140] B. Devincre, R. Madec, G. Monnet, S. Queyreau, R. Gatti, and L. Kubin. Modeling crystal plasticity with dislocation dynamics simulations: The "microMegas" code. In: *Mechanics of Nano-objects* 1:81–100 (2011), 81–100. URL https://api.semanticscholar.org/CorpusID:226221703.

[141] J. Ambrosiano, L. Greengard, and V. Rokhlin. The fast multipole method for gridless particle simulation. In: *Computer Physics Communications* 481:117–125 (1988), 117–125. DOI: https://doi.org/10.1016/0010-4655(88)90029-x.

[142] G. Po, M. S. Mohamed, T. Crosby, C. Erel, A. El-Azab, and N. Ghoniem. Recent progress in discrete dislocation dynamics and its applications to micro plasticity. In: *Jom* 6610:2108–2120 (2014), 2108–2120. DOI: https://doi.org/10.1007/s11837-014-1153-2.

[143] Z. Zhuang, Z. Liu, and Y. Cui. *Dislocation mechanism-based crystal plasticity: theory and computation at the micron and submicron scale*. Academic Press, 2019. ISBN: 978-0128145913. DOI: https://doi.org/10.1016/C2017-0-01936-5.

[144] J. F. Nye. Some geometrical relations in dislocated crystals. In: *Acta metallurgica* 12:153–162 (1953), 153–162. DOI: https://doi.org/10.1016/0001-6160(53)90054-6.

[145] E. Kröner. Berechnung der elastischen Konstanten des Vielkristalls aus den Konstanten des Einkristalls. In: *Zeitschrift für Physik* 1514:504–518 (1958), 504–518. DOI: https://doi.org/10.1007/BF01337948.

[146] E. Kröner. Elasticity theory of materials with long range cohesive forces. In: *International Journal of Solids and Structures* 35:731–742 (1967), 731–742. DOI: https://doi.org/10.1016/0020-7683(67)90049-2.

[147] J. F. Nye. *Physical properties of crystals: their representation by tensors and matrices*. Oxford university press, 1985. ISBN: 0198511655.

[148] A. El-Azab. Statistical mechanics treatment of the evolution of dislocation distributions in single crystals. In: *Physical Review B* 6118:11956 (2000), 11956. DOI: https://doi.org/10.1103/physrevb.61.11956.

[149] T. Hochrainer, M. Zaiser, and P. Gumbsch. A three-dimensional continuum theory of dislocation systems: kinematics and mean-field formulation. In: *Philosophical Magazine* 878-9:1261–1282 (2007), 1261–1282. DOI: https://doi.org/10.1080/14786430600930218.

[150] S. Sandfeld, T. Hochrainer, P. Gumbsch, and M. Zaiser. Numerical implementation of a 3D continuum theory of dislocation dynamics and application to micro-bending. In: *Philosophical Magazine* 9027-28:3697–3728 (2010), 3697–3728. DOI: https://doi.org/10.1080/14786430903236073.

[151] S. Sandfeld, T. Hochrainer, M. Zaiser, and P. Gumbsch. Continuum modeling of dislocation plasticity: Theory, numerical implementation, and validation by discrete dislocation simulations. In: *Journal of Materials Research* 265:623–632 (2011), 623–632. DOI: https://doi.org/10.1557/jmr.2010.92.

[152] K. Schulz, D. Dickel, S. Schmitt, S. Sandfeld, D. Weygand, and P. Gumbsch. Analysis of dislocation pile-ups using a dislocation-based continuum theory. In: *Modelling and Simulation in Materials Science and Engineering* 222:025008 (2014), 025008. DOI: https://doi.org/10.1088/0965-0393/22/2/025008.

[153] M. Monavari, S. Sandfeld, and M. Zaiser. Continuum representation of systems of dislocation lines: A general method for deriving closed-form evolution equations. In: *Journal of the Mechanics and Physics of Solids* 95:575–601 (2016), 575–601. DOI: https://doi.org/10.1016/j.jmps.2016.05.009.

[154] A. M. Kosevich. Dynamical theory of dislocations. In: *Soviet Physics Uspekhi* 76:837 (1965), 837. DOI: https://doi.org/10.1070/pu1965v007n06abeh003688.

[155] I. Groma. Link between the microscopic and mesoscopic length-scale description of the collective behavior of dislocations. In: *Physical Review B* 5610:5807 (1997), 5807. DOI: https://doi.org/10.1103/physrevb.56.5807.

[156] M. Zaiser, M.-C. Miguel, and I. Groma. Statistical dynamics of dislocation systems: The influence of dislocation-dislocation correlations. In: *Physical Review B* 6422:224102 (2001), 224102. DOI: https://doi.org/10.1103/PhysRevB.64.224102.

[157] I. Groma, F. Csikor, and M. Zaiser. Spatial correlations and higher-order gradient terms in a continuum description of dislocation dynamics. In: *Acta Materialia* 515:1271–1281 (2003), 1271–1281. DOI: https://doi.org/10.1016/s1359-6454(02)00517-7.

[158] J. Kratochvíl and R. Sedláček. Statistical foundation of continuum dislocation plasticity. In: *Physical Review B—Condensed Matter and Materials Physics* 7713:134102 (2008), 134102. DOI: https://doi.org/10.1103/PhysRevB.77.134102.

[159] G. Po and N. Ghoniem. A variational formulation of constrained dislocation dynamics coupled with heat and vacancy diffusion. In: *Journal of the Mechanics and Physics of Solids* 66:103–116 (2014), 103–116. DOI: https://doi.org/10.1016/j.jmps.2014.01.012.

[160] D. Steinberger, R. Gatti, and S. Sandfeld. A universal approach towards computational characterization of dislocation microstructure. In: *JOM* 688:2065–2072 (2016), 2065–2072. DOI: https://doi.org/10.1007/s11837-016-1967-1.

[161] H. Song, N. Gunkelmann, G. Po, and S. Sandfeld. Data-mining of dislocation microstructures: concepts for coarse-graining of internal energies. In: *Modelling and simulation in materials science and engineering* 293:035005 (2021), 035005. DOI: https://doi.org/10.1088/1361-651X/abdc6b.

[162] B. Weger, S. Gupta, and T. Hochrainer. Analysing discrete dislocation data using alignment and curvature tensors. In: *Comptes Rendus. Physique* 22S3:249–266 (2021), 249–266. DOI: https://doi.org/10.5802/crphys.60.

[163] A. Demirci, D. Steinberger, M. Stricker, N. Merkert, D. Weygand, and S. Sandfeld. Statistical analysis of discrete dislocation dynamics simulations: initial structures, cross-slip and microstructure evolution. In: *Modelling and Simulation in Materials Science and Engineering* 317:075003 (2023), 075003. DOI: https://doi.org/10.1088/1361-651x/acea39.

[164] C. Zhang, H. Song, D. Oliveros, A. Fraczkiewicz, M. Legros, and S. Sandfeld. Data-mining of in-situ TEM experiments: On the dynamics of dislocations in CoCrFeMnNi alloys. In: *Acta Materialia* 241:118394 (2022), 118394. DOI: https://doi.org/10.1016/j.actamat.2022.118394.

[165] M. F. Ashby. Overview No. 80: On the engineering properties of materials. In: *Acta metallurgica* 375:1273–1293 (1989), 1273–1293. DOI: https://doi.org/10.1016/0001-6160(89)90158-2.

[166] P. Ramalhete, A. Senos, and C. Aguiar. Digital tools for material selection in product design. In: *Materials & Design (1980-2015)* 315:2275–2287 (2010), 2275–2287. DOI: https://doi.org/10.1016/j.matdes.2009.12.013.

[167] S. Gideon. Estimating the dimension of a model. In: *The annals of statistics* 62:461 (1978), 461. DOI: https://doi.org/10.1214/aos/1176344136.

[168] MatWeb Material Property Data. 2025. URL https://www.matweb.com (visited on 08/26/2025).

[169] G. Friedbacher and H. Fuchs. Classification of scanning probe microscopies. In: *Pure and applied chemistry* 717:1337–1357 (1999), 1337–1357. DOI: https://doi.org/10.1351/pac199971071337.

[170] Z. Chen, S. Hengxu, and S. Stefan. *MulDiLingo*. https://gitlab.com/computational-materials-science/public/publication-data-and-code/2022_Zhang_ActaMater_HEA_TEM_coarse_graining.git. 2022.

[171] K. Wada. *Labelme: image polygonal annotation with Python*. 2016. URL https://github.com/wkentaro/labelme.

[172] C. Zhang, C. Bos, S. Sandfeld, and R. Schwaiger. Unsupervised learning of nanoindentation data to infer microstructural details of complex materials. In: *Frontiers in Materials* 11:1440608 (2024), 1440608. DOI: https://doi.org/10.3389/fmats.2024.1440608.

[173] K. von Klinski-Berger. Charakterisierung von Kupfer-Chrom-Verbundwerkstoffen für die Schalttechnik. PhD thesis. Technische Universität Darmstadt, 2015. URL https://d-nb.info/1072647265/04.

[174] D. A. Reynolds et al. Gaussian mixture models. In: *Encyclopedia of Biometrics*. Springer, 2009, pp. 659–663. ISBN: 978-0387730035. DOI: https://doi.org/10.1007/978-0-387-73003-5_196.

[175] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. In: *IEEE transactions on speech and audio processing* 31:72–83 (1995), 72–83. DOI: https://doi.org/10.1109/89.365379.

[176] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12:2825–2830 (2011), 2825–2830. ISSN: 1532-4435.

[177] D. Öztuna, A. H. Elhan, and E. Tüccar. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. In: *Turkish Journal of Medical Sciences* 363:171–176 (2006), 171–176. URL https://api.semanticscholar.org/CorpusID:73014993.

[178] C. Bos. Micromechanical characterization of heterogeneous materials, statistical analysis of nanoindentation data. PhD thesis. Technische Universität Darmstadt, 2019. DOI: https://doi.org/10.5445/IR/1000098226.

[179] K. T. Jacob, S. Priya, and Y. Waseda. A thermodynamic study of liquid Cu-Cr alloys and metastable liquid immiscibility. In: *International Journal of Materials Research* 917:594–600 (2000), 594–600. DOI: https://doi.org/10.1515/ijmr-2000-910710.

[180] D. J. Chakrabarti and D. E. Laughlin. The Cr-Cu (chromium-copper) system. In: *Bulletin of Alloy Phase Diagrams* 51:59–68 (1984), 59–68. DOI: https://doi.org/10.1007/bf02868727.

[181] A. B. Lebedev, Y. A. Berunkov, A. E. Romanov, V. I. Kopylov, V. P. Filonenko, and V. G. Gryaznov. Softening of the elastic modulus in submicrocrystalline copper. In: *Materials Science and Engineering A* 203:165–170 (1995), 165–170. DOI: https://doi.org/10.1016/0921-5093(95)09868-2.

[182] J. E. Chacón and A. I. Rastrojo. Minimum adjusted Rand index for two clusterings of a given size. In: *Advances in Data Analysis and Classification* 171:125–133 (2023), 125–133. DOI: https://doi.org/10.1007/s11634-022-00491-w.

[183] J. Wang, T. Volz, S. Weygand, and R. Schwaiger. The indentation size effect of single-crystalline tungsten revisited. In: *Journal of Materials Research* 3611:2166–2175 (2021), 2166–2175. DOI: https://doi.org/10.1557/s43578-021-00221-6.

[184] X. Ma, W. Higgins, Z. Liang, D. Zhao, G. M. Pharr, and K. Y. Xie. Exploring the origins of the indentation size effect at submicron scales. In: *Proceedings of the National Academy of Sciences* 11830:e2025657118 (2021), e2025657118. DOI: https://doi.org/10.1073/pnas.2025657118.

APPENDICES

A. Derivation of the Expectation-Maximization Updates

This appendix provides detailed derivations of the maximum likelihood estimation procedure and the expectation-maximization (EM) algorithm for Gaussian mixture models, complementing the summary given in Section 5.3. While the main text presents the resulting update rules for clarity, the following sections show step-by-step how these formulas are obtained from Bayes' rule and the log-likelihood function.

A.1. Derivation of Maximum Likelihood Estimation

The posterior probability (responsibility) that datum x_i belongs to component j follows from Bayes' rule,

$$p(j \mid x_i) = \frac{p(x_i \mid j)\, p(j)}{\sum_{l=1}^{k} p(x_i \mid l)\, p(l)}. \tag{A.1}$$

Here $p(x_i \mid j)$ is the Gaussian density of component j, and $p(j) = \alpha_j$ since α_j is the mixture weight of component j. Thus,

$$p(j \mid x_i) = \frac{\alpha_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)}{\sum_{l=1}^{k} \alpha_l \mathcal{N}(x_i \mid \mu_l, \sigma_l^2)}. \tag{A.2}$$

In the EM algorithm, the E-step treats $p(j \mid x_i)$ as fixed during each iteration. The log-likelihood of the dataset $\mathcal{D} = \{x_1, \ldots, x_N\}$ is then

$$\ln \mathcal{L} = \sum_{i=1}^{N} \ln \left[\sum_{j=1}^{k} \alpha_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2) \right]. \tag{A.3}$$

Taking the derivative of $\ln \mathcal{L}$ with respect to parameter $\theta_j \in \{\mu_j, \sigma_j\}$ gives

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_j} = \sum_{i=1}^{N} \frac{\alpha_j}{\sum_{l=1}^{k} \alpha_l \mathcal{N}(x_i \mid \mu_l, \sigma_l^2)} \left[\frac{\partial}{\partial \theta_j} \mathcal{N}(x_i \mid \mu_j, \sigma_j^2) \right]. \tag{A.4}$$

This can be rearranged as

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_j} = \sum_{i=1}^{N} \left[\frac{\alpha_j \mathcal{N}(x_i \mid \mu_j, \sigma_j^2)}{\sum_{l=1}^{k} \alpha_l \mathcal{N}(x_i \mid \mu_l, \sigma_l^2)} \right] \left[\frac{\partial}{\partial \theta_j} \ln \mathcal{N}(x_i \mid \mu_j, \sigma_j^2) \right]. \tag{A.5}$$

The first term is the responsibility $p(j \mid x_i)$; the second term is the derivative of the log-density,

$$\ln \mathcal{N}(x_i \mid \mu_j, \sigma_j^2) = -\tfrac{1}{2}\ln(2\pi) - \ln \sigma_j - \frac{(x_i - \mu_j)^2}{2\sigma_j^2}. \tag{A.6}$$

Consequently,

$$\frac{\partial \ln \mathcal{L}}{\partial \theta_j} = -\sum_{i=1}^{N} p(j \mid x_i)\, \frac{\partial}{\partial \theta_j}\left[\ln \sigma_j + \frac{(x_i - \mu_j)^2}{2\sigma_j^2}\right]. \tag{A.7}$$

Setting derivatives to zero yields the familiar update formulas,

$$\mu_j = \frac{\sum_{i=1}^{N} p(j \mid x_i)\, x_i}{\sum_{i=1}^{N} p(j \mid x_i)} \tag{A.8}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{N} p(j \mid x_i)\, (x_i - \mu_j)^2}{\sum_{i=1}^{N} p(j \mid x_i)} \tag{A.9}$$

$$\alpha_j = \frac{1}{N}\sum_{i=1}^{N} p(j \mid x_i). \tag{A.10}$$

A.2. Illustrative Iteration of the EM Algorithm

Consider a 1D Gaussian mixture with two components (clusters a and b). The EM algorithm proceeds iteratively as follows:

(1) E-step. Assume current parameters $(\mu_a, \sigma_a^2, \mu_b, \sigma_b^2, \alpha_a, \alpha_b)$ are fixed. The likelihood of x_i under component b is given by

$$p(x_i \mid b) = \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\!\left(-\frac{(x_i - \mu_b)^2}{2\sigma_b^2}\right). \tag{A.11}$$

The posterior responsibility of component b for datum x_i is

$$b_i = p(b \mid x_i) = \frac{p(x_i \mid b)\, \alpha_b}{p(x_i \mid b)\, \alpha_b + p(x_i \mid a)\, \alpha_a}, \tag{A.12}$$

with $a_i = 1 - b_i$.

(2) M-step. Update the parameters as weighted averages,

$$\mu_b = \frac{\sum_{i=1}^{N} b_i x_i}{\sum_{i=1}^{N} b_i}, \qquad \sigma_b^2 = \frac{\sum_{i=1}^{N} b_i (x_i - \mu_b)^2}{\sum_{i=1}^{N} b_i}, \tag{A.13}$$

$$\mu_a = \frac{\sum_{i=1}^{N} a_i x_i}{\sum_{i=1}^{N} a_i}, \qquad \sigma_a^2 = \frac{\sum_{i=1}^{N} a_i (x_i - \mu_a)^2}{\sum_{i=1}^{N} a_i}. \tag{A.14}$$

The priors are updated as

$$\alpha_b = \frac{1}{N}\sum_{i=1}^{N} b_i, \qquad \alpha_a = 1 - \alpha_b. \tag{A.15}$$

Repeating E and M steps yields convergence to a local maximum of the log-likelihood.

B. Supplementary Validation on the S-curve

In addition to the main test cases presented in Chapter 4, we provide here supplementary results on the analytic S-curve to illustrate the numerical robustness of the implementation.These plots serve as consistency checks: they are not essential for the main argument, but they demonstrate the correct representation of fields within the discrete-to-continum framework.

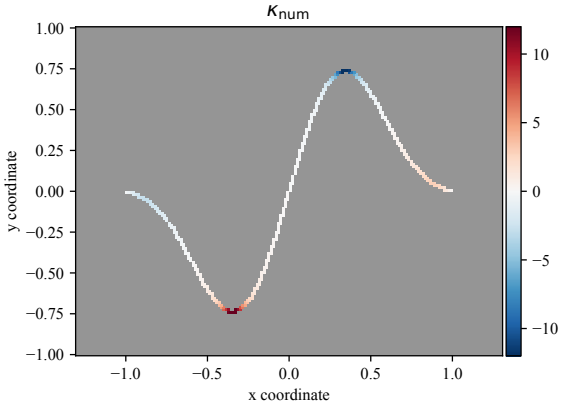

Figure B.1. Comparison between κ_{num} and κ_{cg}, confirming the correctness of κ_{cg}.

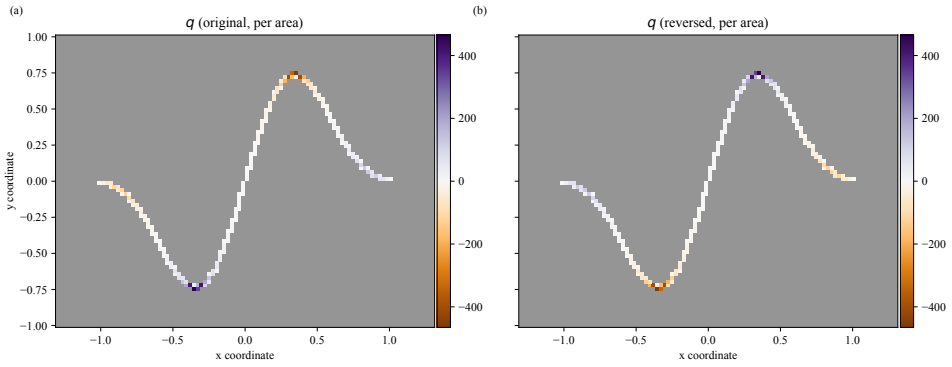

Figure B.2. Orientation-flip validation on the S-curve: upon reversing the parametrization order, the line-length density ρ remains unchanged, while the curvature density q and coarse-grained curvature κ_{cg} correctly change sign, confirming robust handling of signed quantities.

C. Principal Component Analysis in Three Dimensions

The figure of PCA projection in 3D is provided as supplementary material for visualization and reproducibility, rather than for detailed discussion in the main text in Chapter 5.

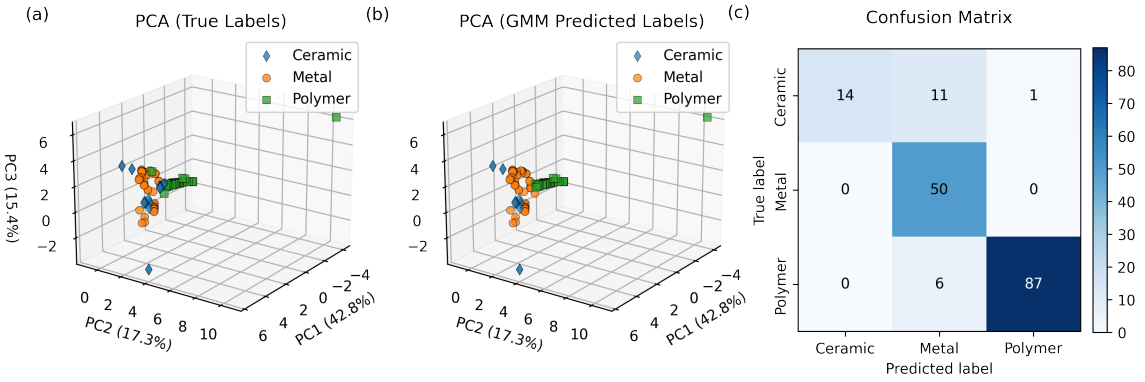

Figure C.1. Clustering results obtained with GMM model m1. (a) PCA projection with true labels in 3D (Metal vs. Polymer vs. Ceramic). (b) PCA projection with GMM-predicted clusters in 3D. (c) Confusion matrix showing 89.35% accuracy.

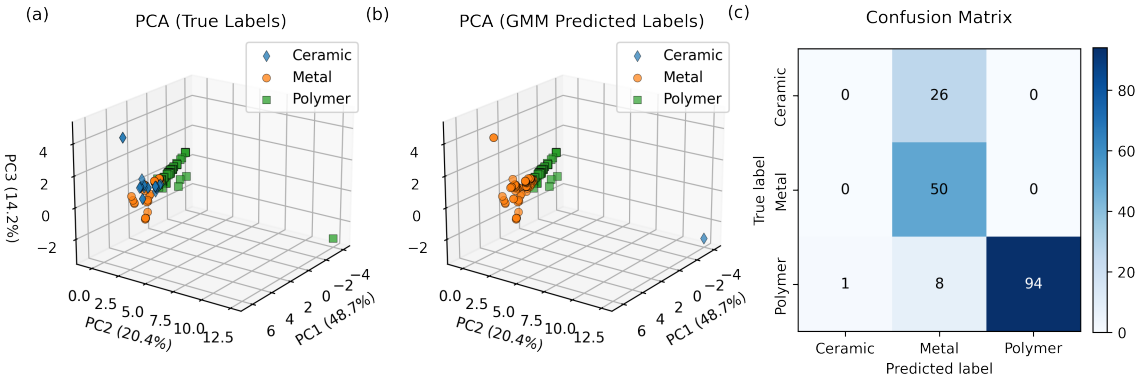

Figure C.2. Clustering results obtained with GMM model m2. (a) PCA projection with true labels in 3D (Metal vs. Polymer vs. Ceramic). (b) PCA projection with GMM-predicted clusters in 3D. (c) Confusion matrix showing 80.45% accuracy.

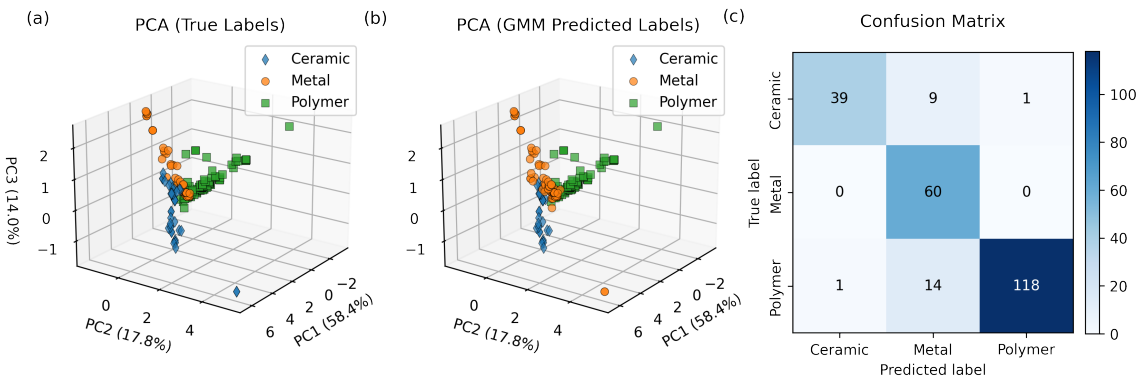

Figure C.3. Clustering results obtained with GMM model m3a. (a) PCA projection with true labels in 3D (Metal vs. Polymer vs. Ceramic). (b) PCA projection with GMM-predicted clusters in 3D. (c) Confusion matrix showing 89.67% accuracy.

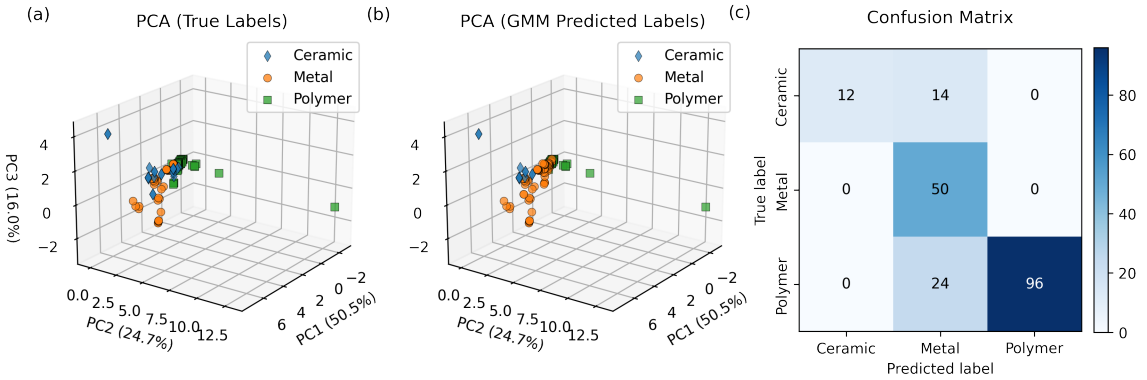

Figure C.4. Clustering results obtained with GMM model m3b. (a) PCA projection with true labels in 3D (Metal vs. Polymer vs. Ceramic). (b) PCA projection with GMM-predicted clusters in 3D. (c) Confusion matrix showing 80.61% accuracy.

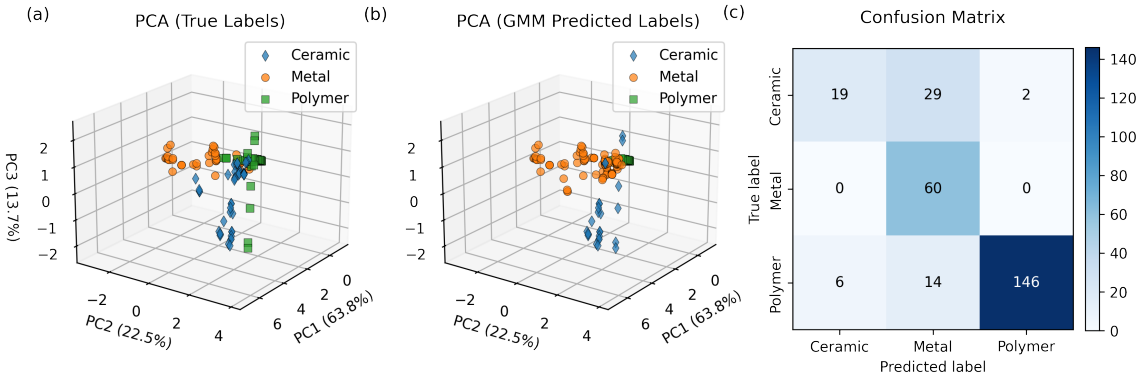

Figure C.5. Clustering results obtained with GMM model m4. (a) PCA projection with true labels in 3D (Metal vs. Polymer vs. Ceramic). (b) PCA projection with GMM-predicted clusters in 3D. (c) Confusion matrix showing 81.52% accuracy.

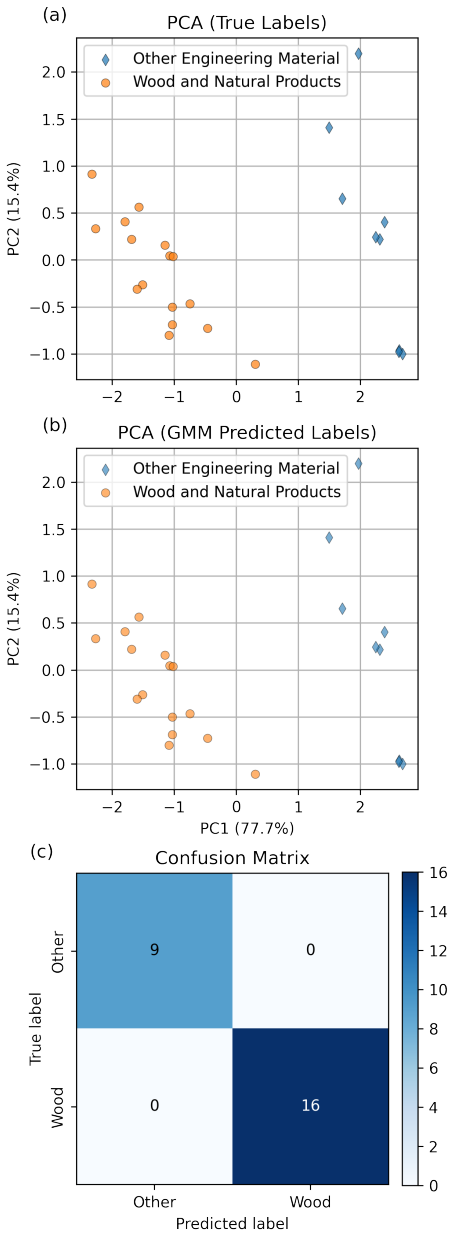

Figure C.6. Clustering results obtained with GMM under a special case. (a) PCA projection with true labels in 2D (Wood vs. Others). (b) PCA projection with GMM-predicted clusters in 2D. (c) Confusion matrix showing 100% accuracy.