# Machine learning workflows for automatic analysis of atom probe tomography data

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

zur Erlangung des akademischen Grades eines

**Doktors der Ingenieurwissenschaften**

genehmigte Dissertation

vorgelegt von

**Alaukik Saxena, M. Sc.**

**Berichter:** Herr Prof. Dr.-Ing. Dierk Raabe
Herr Jun.-Prof. Dr. rer. nat. Benjamin Berkels
Herr Prof. Baptiste Gault, Ph.D.

Tag der mündlichen Prüfung: 07.05.2024

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

# Acknowledgements

I am grateful to Prof. Dierk Raabe for giving me the opportunity to pursue my doctoral research at Max-Planck-Institut für Eisenforschung GmbH. I am thankful to him for helping me understand the bigger picture of my research work, discussions, and his suggestions in my thesis. I am also grateful to Prof. Jörg Neugebauer for many insightful discussions and for giving me the opportunity to do my research in the Computational Materials Design department at Max-Planck-Institut für Eisenforschung GmbH. I am thankful to all my colleagues and supervisors for creating a scientifically vibrant environment that has always motivated me and helped me grow professionally. I am thankful to Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE) and my fellow doctoral students for organizing many data science-related courses and workshops that helped me directly during my PhD.

My gratitude to Dr. Christoph Freysoldt for his support during my PhD journey is difficult to put into words. I am grateful to him for his mentorship, regular supervision, insightful scientific discussions, and, most importantly, his patience. I want to thank him for the corrections and suggestions he made in my thesis. I am thankful to Dr. Baptiste Gault for helping me throughout my PhD through his supervision, interesting scientific discussions, and motivation to think independently. His way of putting complex scientific ideas in a simple way still surprises me and motivates me to approach scientific problems from their fundamentals. I am grateful for the corrections he suggested in my thesis. I am also thankful to Prof. Benjamin Berkels for his supervision, fruitful discussions, and mathematical perspective on my work.

I want to extend my gratitude to Dr. Markus Kühbach and Dr. Sarath Menon for helping me restructure and improve the performance of the workflows I developed during my doctorate. Discussions with them helped me to see my work from the perspective of a software developer. I also want to thank Dr. Paraskevas Kontis, Dr. Alisson Kwiatkowski da Silva, and Dr. Aparna Saksena for providing me with atom probe tomography datasets and many fruitful scientific discussions. I am grateful to Nikita Polin for the experimental datasets, many critical scientific discussions, and fruitful collaboration.

I am grateful to Dr. Shyam Katnagallu for his constant support during my PhD. Regular discussions with him helped me improve my understanding of various complex physical phenomena related to atom probe tomography. I am thankful to him for reviewing parts of my doctoral thesis. I also want to thank Navyanth Kusampudi for critical and stimulating discussions on machine learning.

My PhD journey would never be possible without the emotional support of my friends Sharan Roongta, Navyanth Kusampudi, Aparna Saksena, Qing Tan, Faisal Waqar Syed, Supriya Nandy, and Mahander Pratap Singh.

I am grateful to Perla Sueiras Altamirano for her unconditional support during my PhD. I am grateful to my grandfather, Ishwar Dutt Sharma, for cultivating my interest in science since childhood. I am thankful to my uncle and aunt, Navneet Arora and Seema Arora, for motivating me to pursue higher studies. Finally, I am forever indebted to my parents, Anurag Saxena and Mamta Saxena, and my sister, Chinmayii Saxena, for their lifelong love and support.

# Abstract

The interaction between various microstructural features within a material significantly influences its macroscopic properties. Understanding and measuring the spatial aspects and composition of these microstructural features is crucial for developing new materials and investigating how their structure affects their properties. Atom probe tomography (APT) has emerged as a key technique for material characterization due to its high chemical sensitivity and ability to provide 3D mapping of atomic positions in a material at sub-nanometer resolution. However, analyzing these large datasets (often exceeding 10 million atoms) has traditionally been manual and time-consuming, leading to potential inconsistencies. The integration of machine learning, including clustering and deep learning algorithms, offers a promising solution to streamline this analysis, making it faster and less reliant on manual efforts. In response, this thesis introduces several machine learning-based workflows aimed at automatically quantifying 3D microstructures in APT datasets. These workflows strongly emphasize robustness and reproducibility, holding the potential to drive substantial advancements in the field of materials science.

The initial step in analyzing the 3D microstructure within APT data involves segmenting areas with similar chemical compositions. These regions are termed chemical domains and may correspond to various phases present in a material. The segmentation process is facilitated by a novel, multi-stage unsupervised machine learning approach. To gather local composition information, the APT dataset is divided into uniformly sized voxels. The coordinate system comprising voxel compositions is called composition space. To identify distinct phases or chemical domains within this space, a Gaussian mixture model is employed for clustering. Additionally, a density-based clustering algorithm helps isolate different microstructural features within a single phase at the voxel resolution in real space, treating each feature as a distinct entity. These microstructural entities are then examined for their compositional and geometric properties (such as orientation, shape, and thickness). The methodology, along with its limitations and potential enhancements, is illustrated using both synthetic and actual APT datasets. This includes a detailed examination of a five-component, Fe-doped Sm-Co magnetic alloy and a Nickel-based superalloy that encompasses 30 distinct chemical species.

Quantifying 3D microstructures can be challenging due to their intricate geometries. To address this, a data analysis pipeline using supervised machine learning is developed to examine complex microstructures containing planar subdomains, such as grain boundaries or intertwined plate-like precipitates. The pipeline allows for detailed quantification of segmented planar subdomains, both in terms of their composition (like in-plane composition fluctuations) and their geometry (such as in-plane thickness fluctuations). The robustness and efficiency of this pipeline are exemplified by its successful application to six different APT datasets corresponding to the Fe-doped Sm-Co alloy to find the correlation between material microstructure and magnetic properties.

In order to segment complex linear features, particularly dislocations, a computational geometry concept called skeletonization is used to reduce the iso-composition surface meshes, delineating a microstructure into a linear graph or skeleton. The skeleton effectively encapsulates the topological information corresponding to the 3D iso-composition surface mesh. The underlying skeleton is used to segment dislocations from the meshes for detailed composition analysis. On top of this,

crystallographic information from APT data is used to perform orientation analysis on each dislocation segment to transform it into the crystal coordinate system. This workflow is able to successfully extract and analyze dislocations in a Ni-based alloy and a Fe-Mn alloy, exemplifying its ability to work seamlessly on different APT datasets and its potential to be used as a powerful tool to understand the impact of decorated dislocations on material properties.

Although the 3D APT data is spatially noisy, there are still signatures of the underlying crystal structure of the material in the data. Smooth overlap of atomic orbitals (SOAP) descriptors, which are translationally and rotationally invariant, were used to capture the structural and chemical information around each atom in the APT data. These descriptions are optimized using an autoencoder architecture to accommodate for the noise in the APT data. The optimized descriptors are used to train a neural network to identify ordered $L1_2$ nano-domains in an Al-Mg-Li alloy.

In this thesis, a range of machine learning-based workflows and models are presented, which play a crucial role in deciphering the intricate correlations between material structure and properties. These models facilitate the extraction of descriptors related to the structure and composition of microstructures, laying the ground for training advanced, high-throughput machine learning models.

# Zusammenfassung

Die Wechselwirkung zwischen verschiedenen mikrostrukturellen Merkmalen eines Materials beeinflusst dessen makroskopische Eigenschaften erheblich. Das qualitative Verständnis und die quantitative Beschreibung der geometrischen Aspekte sowie der chemischen Zusammensetzung dieser mikrostrukturellen Merkmale ist entscheidend für die Entwicklung neuer Materialien und die Untersuchung von Struktur-Eigenschafts-Beziehungen. Die Atomsondentomographie (atom probe tomography, APT) hat sich - dank ihrer hohen chemischen Empfindlichkeit und ihrer Fähigkeit, Atompositionen mit einer Auflösung im Subnanometerbereich 3-dimensional zu kartieren, - zu einer Schlüsseltechnik für die Materialcharakterisierung entwickelt. Die Analyse dieser großen Datensätze (oft mehr als 10 Millionen Atome) wird üblicherweise manuell durchgeführt und ist daher zeitaufwändig, was bei der Auswertung über verschiedene Datensätze hinweg zu möglichen Inkonsistenzen führt. Die Integration von maschinellem Lernen, einschließlich Clustering und Deep-Learning-Algorithmen, bietet eine vielversprechende Lösung, um diese Analyse effizienter und schneller zu gestalten, und den manuellen Aufwand zu reduzieren. In dieser Arbeit werden daher mehrere auf maschinellem Lernen basierende Workflows vorgestellt, die auf die automatische Quantifizierung von 3D-Mikrostrukturen in APT-Datensätzen abzielen. Diese Workflows legen besonderen Wert auf Robustheit und Reproduzierbarkeit und haben das Potenzial, wesentliche Fortschritte im Bereich der Materialwissenschaft zu erzielen.

Der erste Schritt bei der Analyse der 3D-Mikrostruktur in APT-Daten ist die Segmentierung von zusammenhängenden Bereichen mit weitgehend konstanter chemischer Zusammensetzung. Diese Regionen werden als chemische Domänen bezeichnet und können z.B. verschiedenen Phasen in einem Material entsprechen. Der Segmentierungsprozess wird durch einen neuartigen, mehrstufigen Ansatz aus dem unüberwachten maschinellen Lernen erleichtert. Um Informationen über die lokale Zusammensetzung zu erhalten, wird der APT-Datensatz in gleich große Voxel unterteilt. Das Koordinatensystem, das die Voxel-Zusammensetzungen umfasst, wird als Kompositionsraum bezeichnet. Zur Identifizierung verschiedener Phasen oder chemischer Domänen innerhalb dieses Raums wird ein Gaußsches Mischungsmodell für die Clusterbildung verwendet. Zusätzlich hilft ein dichtebasierter Clustering-Algorithmus bei der Isolierung verschiedener mikrostruktureller Merkmale innerhalb einer einzelnen Phase mit Voxelauflösung im realen Raum, wobei jedes Merkmal als eigenständige Einheit behandelt wird. Diese mikrostrukturellen Einheiten werden dann auf ihre Zusammensetzung und ihre geometrischen Eigenschaften (wie Orientierung, Form und Dicke) untersucht. Die Methode sowie ihre Grenzen und möglichen Verbesserungen werden anhand von synthetischen und realen APT-Datensätzen veranschaulicht. Das beinhaltet eine detaillierte Untersuchung einer fünfkomponentigen, Fe-dotierten Sm-Co-Magnetlegierung und einer Superlegierung auf Nickelbasis, die 30 verschiedene chemische Spezies umfasst.

Die Quantifizierung von 3D-Mikrostrukturen kann aufgrund ihrer komplizierten Geometrien eine Herausforderung darstellen. Deshalb wurde eine Datenanalyse-Pipeline mit überwachtem maschinellem Lernen entwickelt, um komplexe Mikrostrukturen zu untersuchen, die planare Unterbereiche wie Korngrenzen oder ineinander verschachtelte plattenförmige Ausscheidungen enthalten. Die Pipeline ermöglicht eine detaillierte Quantifizierung segmentierter planarer Subdomänen, sowohl

in Bezug auf ihre Zusammensetzung (z. B. Schwankungen der Zusammensetzung in der Ebene) als auch auf ihre Geometrie (z. B. Schwankungen der Dicke in der Ebene). Die Robustheit und Effizienz dieser Pipeline wird durch ihre erfolgreiche Anwendung auf sechs verschiedene APT-Datensätze der Fe-dotierten Sm-Co-Legierung veranschaulicht, um Korrelationen zwischen der Mikrostruktur des Materials und den magnetischen Eigenschaften zu finden.

Um komplexe lineare Entitäten, insbesondere Versetzungen, zu segmentieren, wird ein Konzept der algorithmischen Geometrie, die so genannte Skelettierung, verwendet, um die Konzentrations-Isoflächen zu reduzieren und eine Mikrostruktur in einen linearen Graphen, ein sogenanntes Skelett, zu verwandeln. Das Skelett abstrahiert die topologische Informationen, die der 3D-Isofläche entspricht. Das so erhaltene Skelett wird verwendet, um einzelne Versetzungen zu segmentieren und damit eine detaillierte Analyse der Konzentrationsprofile zu ermöglichen. Darüber hinaus werden kristallografische Informationen aus APT-Daten verwendet, um eine Orientierungsanalyse für jedes Versetzungssegment durchzuführen und es in das Kristallkoordinatensystem zu transformieren. Dieser Workflow kann - mit Erfolg - Versetzungen sowohl in einer Ni-basierten-Superlegierung als auch in einer Fe-Mn-Legierung extrahieren und nachfolgend zu analysieren. Diesverdeutlicht seine Fähigkeit, mühelos mit unterschiedlichen APT-Datensätzen umzugehen, und zeigt sein Potenzial, als leistungsfähiges Werkzeug zum Verständnis der Rolle von dekorierten Versetzungen für die Materialeigenschaften zu dienen.

Obwohl die 3D-APT-Daten in der Ortsauflösung stark verrauscht sind, lassen sich in den Daten dennoch die Signatur der zugrunde liegenden Kristallstruktur finden. Um die strukturellen und chemischen Informationen in der Umgebung jedes Atoms in den APT-Daten zu erfassen, wurden SOAP-Deskriptoren (Smooth Overlap of Atomic Orbitals) verwendet, die translations- und rotationsinvariant sind. Diese Deskriptoren werden mit einer Autoencoder-Architektur optimiert, um dem Rauschen in den APT-Daten Rechnung zu tragen. Die optimierten Deskriptoren werden verwendet, um ein neuronales Netzwerk zu trainieren, das geordnete $L1_2$-Nanodomänen in einer Al-Mg-Li-Legierung identifiziert.

In dieser Arbeit werden mehrere auf maschinellem Lernen basierende Workflows und Modelle vorgestellt, die entscheidend zur Entschlüsselung der komplizierten Zusammenhänge zwischen Materialstruktur und -eigenschaften beitragen. Die Modelle erleichtern die Extraktion von Deskriptoren für die Struktur und Zusammensetzung der Mikrostruktur, und legen somit den Grundstein, um zukünftig fortschrittlicher Modelle des maschinellen Lernens mit hohem Durchsatz zu trainieren.

# List of Abbreviations

| | |
|---|---|
| APT | Atom Probe Tomography |
| ML | Machine Learning |
| AI | Artificial Intelligence |
| PCA | Principal Component Analysis |
| GMM | Gaussian Mixture Model |
| BIC | Bayesian information criterion |
| DBSCAN | Density-Based Spatial Clustering of Applications with Noise |
| UMAP | Uniform Manifold Approximation and Projection |
| t-SNE | t-distributed Stochastic Neighbor Embedding |
| DT | Decision Tree |
| DNN | Deep Neural Networks |
| ReLU | Rectified Linear Unit (ReLU) |
| RF | Random Forest |
| MNIST | Modified National Institute of Standards and Technology database |
| ROI | Region of Interest |
| MCF | Mean Curvature Flow |
| SOAP | Smooth Overlap of Atomic Positions |
| FFT | Fast Fourier Transform |
| SDM | Spatial Distribution Map |
| FCC | Face Centered Cubic |
| TEM | Transmission Electron Microscopy |
| HAADF | High Angle Annular Dark Field imaging |
| HRTEM | High Resolution Transmission Electron Microscopy |

# Contents

**Acknowledgements**

**Abstract**

**Zusammenfassung**

**List of Abbreviations**

**1 Introduction** **1**
1.1 Motivation . . . 1
1.2 Thesis outline . . . 6

**2 Machine learning and data analysis methods** **8**
2.1 Supervised learning . . . 8
Random forests . . . 9
Deep Neural Networks . . . 11
U-Net . . . 12
2.2 Unsupervised learning . . . 14
2.2.1 Clustering . . . 14
Gaussian mixture models . . . 14
Density-based spatial clustering of applications with noise (DBSCAN) . . . 15
2.2.2 Dimensionality reduction . . . 17
Principle component analysis (PCA) . . . 17

**3 Chemical domains segmentation: Clustering in composition space-based approach** **19**
3.1 Introduction . . . 19
3.2 Materials and Methods . . . 20
3.2.1 APT Data . . . 20
Experimental Data . . . 20
Synthetic data . . . 21
3.2.2 Classification of chemical domains: clustering in composition space . . . 21
3.3 Results and Discussion . . . 24
3.3.1 Visual validation with isosurface results: . . . 24
3.3.2 Calculation of volume fraction and composition of each phase: . . . 24
3.3.3 Quantitative comparison with isosurfaces . . . 26
3.3.4 Dimensionality reduction for composition space . . . 28
3.4 Conclusion . . . 33

**4 Machine learning based decomposition of geometrically complex chemical domains 34**
4.1 Introduction . . . 34
4.2 Materials and Methods . . . 35
4.2.1 Experimental data . . . 35
4.2.2 APT data analysis pipeline . . . 35
Decomposing geometrically complex precipitates . . . 36
4.3 Results and discussion . . . 39
4.3.1 In-plane composition and thickness fluctuations . . . 42
4.3.2 Quantification of the microstructure . . . 48
4.4 Conclusion . . . 49

**5 Analyzing linear features in atom probe tomography datasets using skeletonization 51**
5.1 Introduction . . . 51
5.2 Materials and Methods . . . 52
5.2.1 Experimental Data . . . 52
5.2.2 Workflow . . . 53
5.2.3 Curve skeleton extraction . . . 53
5.2.4 Distance from a B-spline . . . 55
5.2.5 Validation on synthetic data . . . 55
5.3 Results and discussion . . . 58
5.3.1 Synthetic data . . . 58
5.3.2 Morphology based segmentation . . . 59
5.3.3 Orientation analysis . . . 63
Orientation analysis algorithm . . . 65
Application on Fe-Mn alloy . . . 67
5.4 Conclusion . . . 68

**6 Applying supervised learning to extract local ordering information from APT data 73**
6.1 Introduction . . . 73
6.2 Methods and materials . . . 74
6.2.1 Experimental data . . . 74
6.2.2 Smooth Overlap of Atomic Positions (SOAP) . . . 74
6.2.3 Learning local atomic environments in APT . . . 75
6.2.4 Dimensionality reduction and hyperparameter tuning using autoencoders . . . 76
6.3 Results and discussion . . . 78
6.3.1 Autoencoder application to the experimental data . . . 80
6.3.2 Neural network application on the experimental data . . . 82
6.4 Conclusion . . . 84

**7 Summary and Outlook 85**
7.1 Outlook . . . 87

**Copyright 90**

**Bibliography 91**

**List of Figures 106**

**List of Tables 107**

# Chapter 1

# Introduction

## 1.1 Motivation

New materials have historically propelled significant advancements in civilization, such as bronze aiding agriculture or semiconductors enabling modern computing. Yet their production poses a severe environmental challenge as it is one of the primary source of greenhouse gas emissions, energy usage, and pollution. This necessitates a radical re-evaluation of how we produce, utilize, and recycle materials to mitigate environmental impacts while continuing to enable innovation [1], [2]. Material design typically entails modifying the composition and processing conditions, guided by systematically planned experiments or computer simulations to create new materials [1], [3], [4]. However, the vast chemical space and intricate material processing-property relationships make the task of designing new materials via conventional approaches challenging. Consequently, there has been a recent surge in machine learning-guided materials design methods. Usually, understanding the processing-structure-property relationships of a particular material by supervised machine learning requires collecting enough experimental or simulation training data. Following this, descriptors or features that encapsulate the relevant information in the data are extracted [5]. The task of extracting descriptors from training data varies in difficulty. For example, it is simpler with two-dimensional (2D) microstructure images where relevant features can be labeled by humans but more complex when describing 3D atomic environments in atomic-scale simulations. In this thesis, I have formulated automated workflows to discern and extract descriptors from geometrically complex 3D experimental microstructures using machine learning. These workflows can be employed across a multitude of 3D experimental datasets, facilitating the training of high-throughput machine learning models to elucidate the relationships between material structure and properties.

The interplay among diverse microstructural features significantly influences the macroscopic properties of bulk materials [6]. Thus, delineating the spatial distribution and compositional attributes of these features is crucial for both tailoring novel materials and elucidating the relationships between properties and structures. Atom probe tomography (APT) has emerged as an important material characterization technique, complementing the capabilities of electron and X-ray-based microscopy techniques [7]. APT can provide a 3D distribution of atoms along with their chemical identities at sub-nanometer resolution within a material [7]–[10]. It facilitates the mapping of various microstructural features, including interface and interphase boundaries, dislocations, and secondary phase regions in 3D space via their compositional fingerprint [11]–[13]. This in-depth spatial and compositional analysis is instrumental in decoding the intricate processing-structure-property relationships.

APT harnesses a strong electric field (around $10^{10}\ Vm^{-1}$) to desorb and ionize atoms from a needle-shaped specimen's surface, a process termed as field evaporation (as illustrated in Figure 1.1(A)) [14]. The necessary evaporation field is established through a constant DC high voltage (HV), with added pulses of HV (or laser pulses for thermal activation). This standing electric field propels the

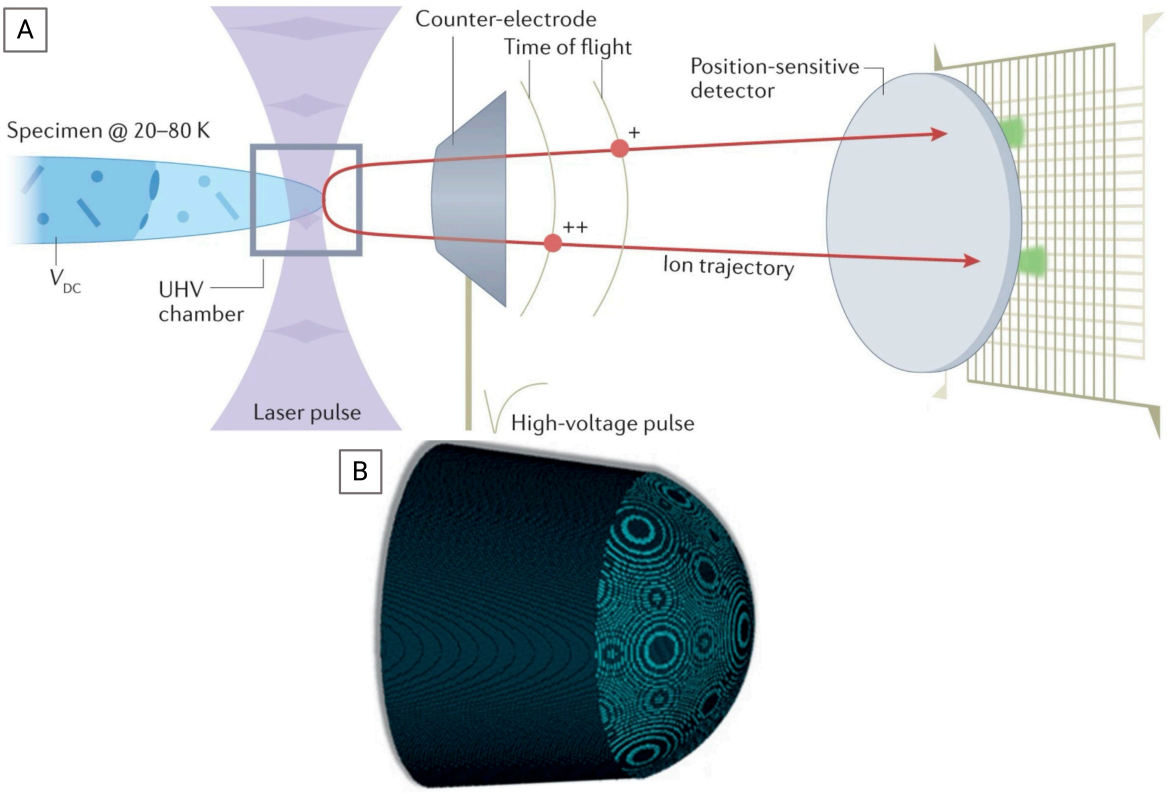



FIGURE 1.1: **(A)** Schematic of an atom probe ultra-high vacuum (UHV) analysis chamber experimental setup. The needle-shaped specimen is biased with a direct high voltage $V_{DC}$, and laser or high-voltage pulses are applied to field evaporate atoms from the surface of the specimen. The ions are accelerated by the standing electric field to hit a position sensitive detector. The time of flight information of each ion is recorded to deduce its chemical identity. (Reproduced with permission from [7]. Copyright © (2021) Springer Nature).**(B)** Schematic of a needle-shaped specimen. The atoms that protrude most at the surface of the specimen are depicted in light blue color. The patterns formed by these atoms on the surface are related to the underlying crystal structure of the specimen. (Reproduced with permission from [8]. Copyright © (2012) Elsevier.)

ions towards a 2D position-sensitive detector. Both the impact position on the detector and each ion's flight time are recorded during the experiment. The specimen's needle-like shape, having a high radius of curvature (20 to 100 nm), creates a significantly divergent electric field around it. This diverging field magnifies the ions' projection on the detector by about a million times, aiding in the direct visualization of ions at the detector. The technique of time-of-flight mass spectrometry is used to reveal the chemical identity of each recorded ion. By analyzing each ion's impact position on the detector and the sequence of hits, the atoms' original positions in the specimen are deduced, forming a reconstructed point cloud [8], [15].

The size of the APT datasets has consistently expanded over the years owing to a wider field-of-view [16] and the incorporation of laser pulsing features [17], [18]. Nowadays, APT datasets commonly encompass tens to hundreds of millions of ions. The data retrieved during the APT experiment includes x-y detector impact coordinates, time of flight, standing voltage and pulsing voltage for each recorded ion. As mentioned above, the detector hit coordinates and sequence of hits are used to calculate the x-y-z coordinates of each atom in the specimen using a reconstruction algorithm. As a post-processing step, the time of flight data of the ions is converted into mass-to-charge state ratios. It is assumed that the potential energy of a field evaporated ion ($neV$) is entirely converted into kinetic energy ($\frac{mv^2}{2}$), where $n$ is the unit charge on the ion, $e$ is the elementary charge of an electron, $V$ is the applied voltage, $m$ is the mass of the ion and $v$ is its velocity. The velocity $v$ is computed as the ratio of $d$ to $t$, where $d$ is the distance between the specimen and the detector, and $t$ is the time-of-flight of the ion. Subsequently, a histogram of mass-to-charge state ratios is computed. Traditionally, an experimentalist assigns each peak and, thereby, each ion corresponding to that peak in the histogram to a chemical species [19]. Following this step, a 3D point cloud with x-y-z coordinates and the chemical identity of each atom is obtained. The reconstructed point cloud corresponding to an aluminium alloy is shown in Figure 1.2(A), where different colors depict the chemical identity of each atom.

Following the detailed 3D atom-by-atom reconstruction of raw APT data, various postprocessing methods are applied to visualize and analyze multiple microstructural features. Broadly, these techniques are classified into two core categories: voxel-based analysis and interatomic distance-based analysis [7].

**Voxel-based analysis**:

In this method, APT reconstruction data is divided into an array of voxels, each either encapsulating a fixed volume or containing a set number of atoms. The composition (molar fraction) or the number density of each voxel is calculated. Isosurfaces, the most widely used technique to outline microstructural features in APT, are then constructed by defining a threshold molar fraction or number density for a specific element (or multiple elements) [20]. A preset molar fraction threshold (expressed in atomic fraction) leads to an iso-composition surface circumscribing the regions with a higher molar fraction surpassing the threshold. Likewise, a defined number density threshold (expressed in the number of atoms per $nm^3$) leads to an iso-density surface. Through careful threshold adjustment, it is feasible to demarcate interfaces between microstructural regions and conduct targeted analysis on these delineated regions as shown in Figure 1.2 (C and D) for a Ni-based alloy. For instance, isosurfaces can unveil the interface between secondary phase precipitates and the encompassing matrix. Further, proximity histograms can be computed to trace the average evolution of composition along the local normals to the isosurfaces [21], [22] as illustrated in Figure 1.2 (E). This tool is crucial for grasping composition fluctuations across interfaces, even in scenarios where the interfaces are disconnected or curved.

**Interatomic distance-based analysis**:

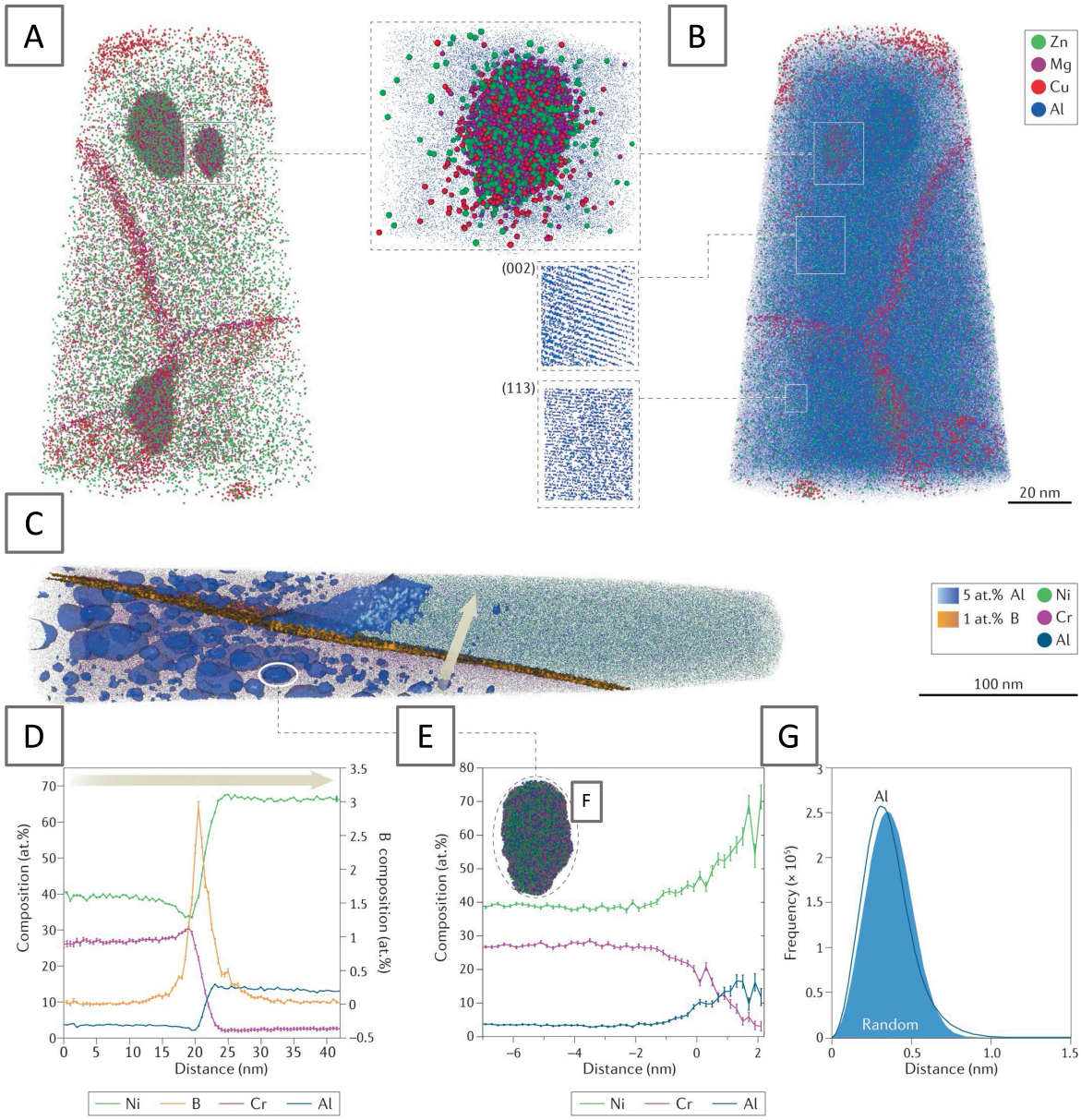



FIGURE 1.2: **(A)** Reconstructed point cloud corresponding to an Al alloy where different colors correspond to different chemical species. Grain boundary segregation and precipitates (one of them also magnified) can be seen. **(B)** Different families of atomic planes in the reconstructed data. **(C)** Reconstructed point cloud corresponding to a Ni-based alloy with the interfaces between $\gamma$ and $\gamma'$ phase and grain boundary segregation delineated by blue and golden iso-composition surfaces, respectively. **(D)** 1D composition profile along the arrow shown in panel (C) across the grain boundary (golden). **(E)** Proximity histogram for the isosurface highlighting the selected precipitate shown in panel **(F)**. **(G)** The first nearest-neighbor distribution of Al atoms deviates from the corresponding random distribution, showing the clustering tendency. (Reproduced with permission from [7]. Copyright © (2021) Springer Nature.)

In this method, the local neighborhood around each atom in the reconstruction data is analyzed by directly measuring distances between the atoms. One straightforward technique is the nearest neighbor analysis, which involves studying the distribution of distances between a solute atom and its neighbors [23]. This distribution is compared to the corresponding one calculated for a system with randomly assigned element labels. This comparison aids in identifying signs of nanoscale clustering. For example, Figure 1.2 (G) illustrates the deviation of the first nearest neighbor distribution of Al in a Ni-based alloy from the corresponding random distribution, showing the tendency of clustering or precipitation. A more widely used method in the present context is the "maximum separation" algorithm [24]. This method is based on the assumption that the distance between a pair of nearest neighbor solute atoms in a cluster is less than a similar pair in the matrix. The maximum distance criteria is used to cluster the solute atoms within the reconstruction data. Further, each cluster can be studied for composition, shape, and size [25], [26]. A statistical analysis can also be performed on the identified clusters to measure features like the average distance between neighboring clusters or the average number of clusters in a given space and analyze how these features relate to the material properties. In recent years, several other algorithms based on different cluster definitions have been developed, for example, density-based clustering algorithms [27] and Gaussian mixture models [28].

While analyzing crystalline materials via APT, due to the needle shape of the specimen, atoms corresponding to sets of low-index crystallographic plane terraces protrude out at the surface and form concentric rings around the crystallographic directions [8] (as shown in Figure 1.1 (B)). These protruding atoms are the first to field evaporate due to the high electric field at the edges of atomic planes [29]. The variation of the electric field at the surface of the specimen causes the field evaporated ion to deviate sideways and leads to trajectory aberrations in APT [14], [30]. These aberrations create areas with different ion hit intensities on the detector, which can be linked to the crystallographic information of the material [31]. The regions of low intensity that appear on the detector in this way are commonly referred to as poles. Since, in most cases the crystal structure of the material being analyzed is known, a scientist compares the patterns formed on the detector hit map to the stereographic projection of the crystal. Thereby, each pole is assigned to a crystallographic direction or plane. The resolution in APT reconstruction data is good enough to distinguish atomic planes near the pole but not enough to resolve the crystal lattice in the lateral direction to the pole [32]. Different families of atomic planes are revealed in the reconstructed point cloud for an Al alloy, shown in Figure 1.2(B). Various methods: Fourier transform-based [33], Hough transforms (an algorithm to capture linear features) [34], and radial distribution-based (spatial distribution maps (SDMs)) [35] have been developed to extract and analyze the atomic planes corresponding to a pole or crystallographic direction from the reconstruction data. Interatomic plane distances, angles between various crystallographic directions, and orientation of various microstructural features relative to the crystal coordinate system can be computed using the above-stated methods. The calculated inter-planar distances have also been used to calibrate the reconstruction data [36].

APT provides chemically resolved 3D distribution of atoms in a material to visualize and study various microstructural features. Application of APT is not just limited to alloy design but encompasses geological and even biological materials [37], [38]. However, the analysis of atom probe data by the above mentioned techniques is still largely performed manually [39]. For example, the molar fraction threshold for constructing isosurfaces or input hyperparameters for clustering algorithms have to be provided and generally manually tuned by the user. Given the size (tens to hundreds of millions of atoms) and complexity of APT data sets this process of analysis, especially while analyzing several datasets, might have questionable quantitative correctness and reproducibility [7]. To this

end, some machine learning (ML) tools have been recently developed to automate APT data analysis. Specifically, deep learning-based methods have been developed to distinguish the interfaces between the matrix and the precipitate phase automatically to replace the use of iso-composition surfaces [40], to autonomously label the poles on the detector hit maps [41], or to assign chemical species to peaks in time-of-flight mass spectrum [42]. Furthermore, convolution neural networks (CNN) based models have been developed to distinguish local ordered domains from the disordered domains utilizing the crystallographic information from APT [43], [44]. There are also ML-based methods to segment out grain boundaries in an automatic way from APT data [45], [46]. These methods push the limits of human-guided atom probe analysis by providing an automated and reliable alternative. They solve the problem for which they were developed but lack generalization, which is a prerequisite while dealing with noisy and complex APT data. For example, the model developed for indexing poles on a detector hit map was trained only on datasets corresponding to Al, which has a clear contrast between the poles and the surrounding regions, which is not very common if multi-species alloys are considered. Additionally, most of these techniques do not provide an automated solution for the most common and time-consuming APT data analysis problem: quantifying complex microstructures. Specific drawbacks of these methods relevant to this thesis are discussed in the following chapters.

Today, more than 50 groups in the world have APT experimental facilities which means that continuously developing better automated ML workflows will not just help individual users to save time but also give an opportunity to do high-throughput ML studies because so much data is produced. This thesis aims to develop automated ML-based workflows that can enable a quantitative description of 3D microstructures in APT data with a focus on reproducibility. Reproducibility is enforced in this study by making sure a systematic workflow development and data management. Most of the developed workflows are already wrapped inside Pyiron [47], an integrated development environment enabling computational materials science tools, making them easy to use and ensuring continuous further development. The key features of this thesis are:

- Developed general unsupervised ML-based workflow to segment out key microstructural features from APT data automatically.
- Developed supervised ML-based workflows to simplify and analyze complex microstructural features containing planar regions like grain boundaries.
- Developed computational geometry tools to segment and analyze complex microstructural features containing linear features like dislocations.
- Extract crystallographic directions in APT data automatically.
- Use the general description of atomic environments to autonomously capture ordered crystal domains in APT data using supervised learning.

## 1.2 Thesis outline

Chapter 2 **Machine learning and data analysis methods**
This chapter explains the general ML and data analysis methods used in this thesis.

Chapter 3 **Chemical domains segmentation: Clustering in composition space-based approach**
This chapter introduces a novel workflow based on unsupervised learning to segment out microstructural features using their chemical fingerprint in the reconstructed APT data. This chapter is a modified version of my publication in Microscopy and Microanalysis entitled

"A machine learning framework for quantifying chemical segregation and microstructural features in atom probe tomography data" [48].

Chapter 4 **Machine learning based decomposition of geometrically complex chemical domains**
This chapter introduces a U-Net architecture-based workflow that I developed to reduce geometrically complex microstructural features extracted by the clustering in the composition space-based approach into simpler plate-like features. Examples of complex microstructural features may include a network of grain boundaries or entangled plate-like precipitates. The workflow provides an in-depth compositional and morphology quantification of the microstructural features. This chapter is a modified version of my publication in Microscopy and Microanalysis entitled "A machine learning framework for quantifying chemical segregation and microstructural features in atom probe tomography data" [48].

Chapter 5 **Analyzing linear features in atom probe tomography datasets using skeletonization**
This chapter uses a topology analysis concept called skeletonization to extract and segment linear features from isosurface meshes in the reconstruction data. The chapter focuses explicitly on segmenting dislocations in a complex microstructure, enabling a thorough composition and orientation analysis of each dislocation. This chapter is a modified version of my publication in Microscopy and Microanalysis entitled "Analyzing linear features in atom probe tomography datasets using skeletonization" [49].

Chapter 6 **Applying supervised learning to extract local ordering information from APT data**
In this chapter, general descriptors based on radial basis functions and spherical harmonics are used to encapsulate local chemical and structural information around each atom in the reconstructed APT data. These descriptors are employed in conjunction with a trained supervised ML model to extract $L1_2$ ordered domains.

# Chapter 2

# Machine learning and data analysis methods

This chapter discusses general machine learning (ML) and data analysis methods used in the latter part of this thesis. Given the abundance of existing literature detailing the mathematical aspects of ML models, this chapter focuses on providing an intuitive understanding of the rationale and concepts underlying these models [50], [51]. All the algorithms and workflows that are built on these ML models are discussed in detail in the subsequent chapters.

Machine learning comprises computer algorithms that can make predictions based on data corresponding to a task without being explicitly modeled or programmed. A more thorough definition of ML was given by Tom Mitchell - *"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E"* [52]. The task $T$ here can be, for example, classifying emails as spam or not spam based on the contents of a given email. The experience, $E$, using which the algorithm learns is the training data, a set of instances in which all the emails are already labeled as spam or not. The performance of the algorithm, $P$, can be evaluated by calculating the number of instances where the algorithm made a correct prediction per total number of instances [53].

Depending on whether labeled data is required to train a model, ML models can be divided into two broad categories: supervised learning and unsupervised learning. In supervised learning, the training data is labeled, which means the input data is already coupled with the correct output, and the model tries to learn the relationship between the input and the output data [50]. On the contrary, in unsupervised learning the training data is not labeled, and the model tries to find patterns in the input data; for example, clustering the data points lying in close proximity to each other by measuring Euclidean distances between them. Most of the machine learning algorithms used in this thesis fall into either of these categories and are discussed below.

## 2.1 Supervised learning

In supervised learning, the training data consists of input data $\mathbf{X} \in \mathbb{R}^{p \times n}$ where $p$ is the total number of training examples, and $n$ is the number of features, and output data $\mathbf{Y}$. If $\mathbf{Y}$ takes real and continuous values, then the training task is called regression; on the other hand, if it takes discrete values, for example, for the task of classifying emails as spam or not spam, then the training task is called classification [51]. The $i$th training input example is denoted as $x^{(i)}$, output by $y^{(i)}$ and the model prediction is denoted as $\hat{y}^{(i)}$. The supervised training model learns the relationship between $\mathbf{X}$ and $\mathbf{Y}$. To enhance the performance of the model, the model parameters are tweaked while minimizing the

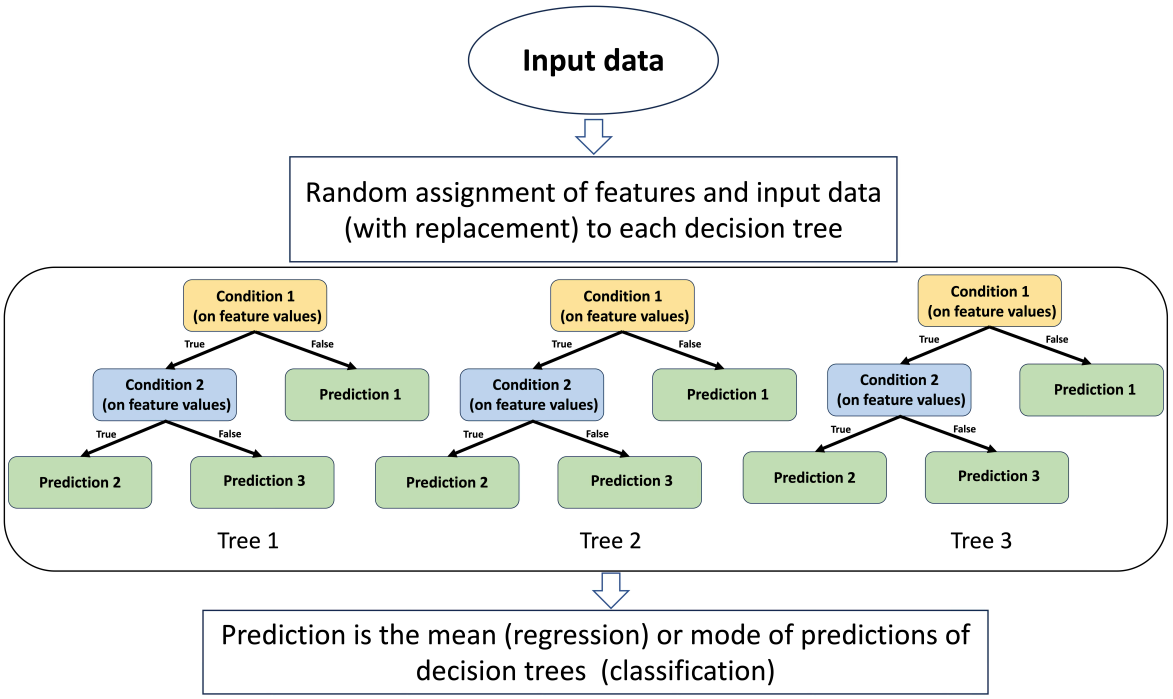



FIGURE 2.1: A schematic of a random forest (RF) model comprising of three decision trees.

error (between the predicted and the actual output) or loss function. For the regression task, squared error (SE) can be the loss function (for a single data point):

$$SE = (y^{(i)} - \hat{y}^{(i)})^2 \tag{2.1}$$

For classification, cross-entropy (CE) is generally chosen as the loss function (for a single data point):

$$CE = -\sum_{c=1}^{C} y_c^{(i)} \log(\hat{y}_c^{(i)}) \tag{2.2}$$

where $C$ is the total number of classes in a given classification task. For example, in the spam classification problem $C = 2$ as there are two classes: spam and not spam. Supervised learning algorithms used in this thesis are discussed below [50].

### Random forests

Random forests (RF) is a supervised learning algorithm and can be used for both classification and regression [54]. An ensemble of decision trees forms a random forest illustrated in Figure 2.1 [50], [54]. A decision tree, another supervised learning model, can also be used for regression and classification tasks [52]. Each tree in the random forest is trained on a random subset of the training data with a random subset of features selected from the given features in the original training data. The random subset of the training data for each tree is sampled with replacement from the original training data. For the classification problem, the mode of all the classes predicted by the decision trees contained in an RF is considered the prediction of the RF model. However, for the regression task, the mean of all predictions by the decision trees is the output prediction of the RF model, making the model more robust and immune to overfitting.

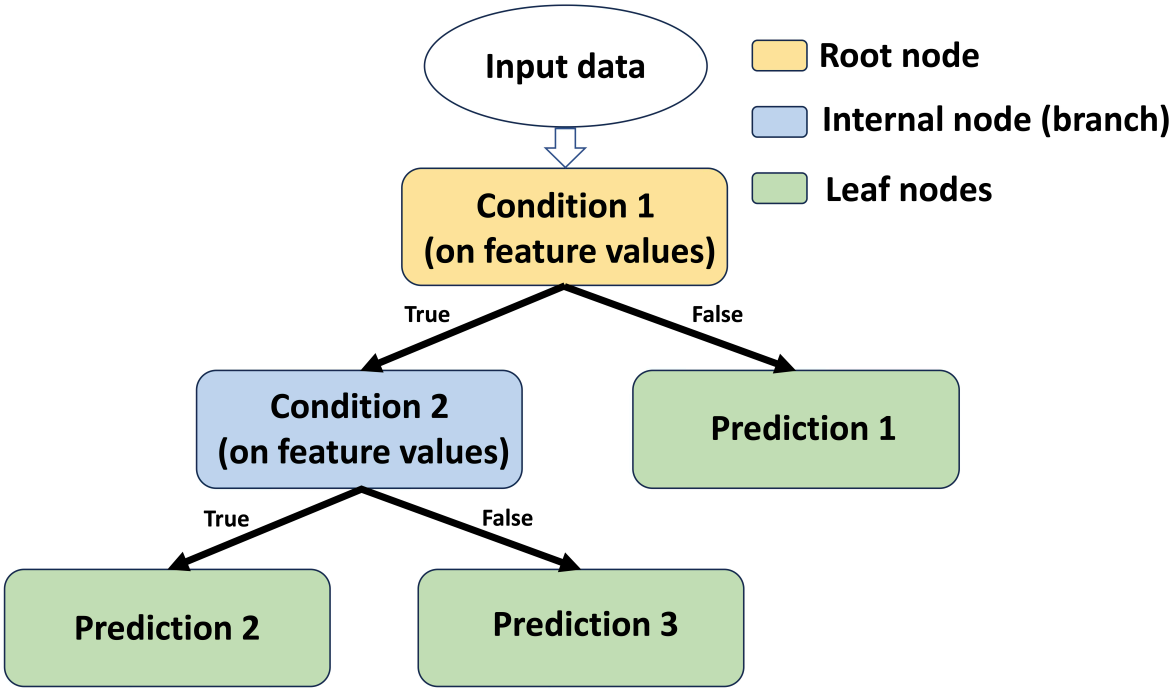



FIGURE 2.2: A schematic of a decision tree (DT) with conditions applied to features in the input data at the root and branch nodes. The predictions are made by the leaf nodes.

Since decision trees are the building blocks of a RF model it is important to understand their working [55]. A decision tree (DT) divides the input data into subsets based on conditions applied to feature values. The conditions that are learnt after training a DT model help to predict classes or values for previously unseen data. A DT model consists of nodes where conditions on features are applied to partition the training data and leaf nodes where all the final predictions are made. The leaf nodes are shown by green color boxes in Figure 2.2. No further division of data into smaller subsets is possible after the leaf nodes. The top node of the tree is called the root node, and the internal nodes are called branches as shown in Figure 2.2. The construction of a decision tree consists of assigning conditions on feature values at each node and deciding whether to split data at a given node. To determine what feature condition will serve as root node or branches and what will be the final structure of the tree, different strategies exist depending on the type of supervised learning task. For a classification task, the Gini impurity or cross-entropy is used to measure the randomness in data at a node [50]. The Gini impurity at a node is given as,

$$\text{Gini} = 1 - \sum_{c=1}^{C} P_c^2 \tag{2.3}$$

where $C$ is the total number of classes in a classification task. $P_c$ is the probability of a class at the given node. The feature condition that has the minimum Gini impurity forms the root node. Also, the feature conditions selected at the subsequent branches are the ones that minimize the Gini impurity. However, for a regression task, the sum of squared residuals at each node is calculated, and the feature conditions that minimize this sum form the root and the branch nodes. After splitting the data into subsets following the local minimization of Gini index or sum of squared residuals at each node, the conditions applied to the feature values are recorded [50]. These feature conditions then guide the model for making predictions for previously unseen data. The feature conditions in the RF model can be easily interpreted and can also be used to calculate the importance of the features in a given input data. In Chapter 3, an RF model is used to calculate the relative importance of different chemical

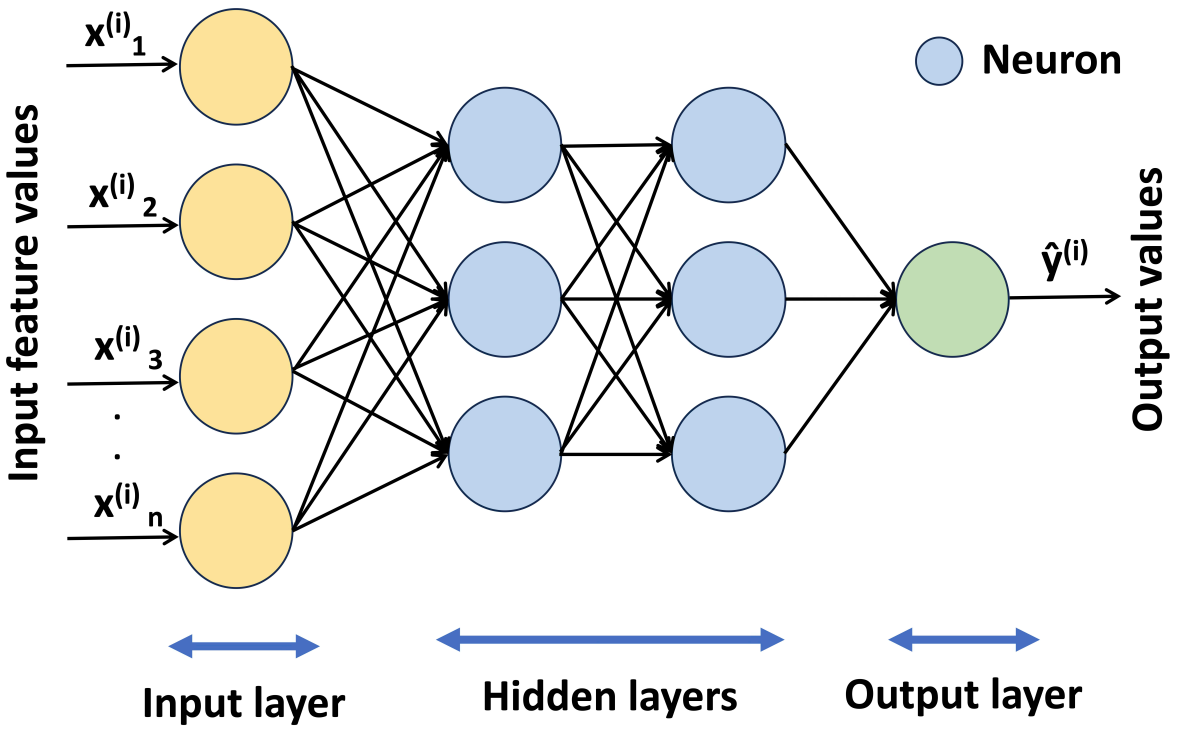



FIGURE 2.3: A schematic of a deep neural network (DNN). The feature values corresponding to the $i^{th}$ training example, $x^{(i)}$, are fed into the input layer. Here, the feature space has $n$ dimensions. The arrows depict the interconnections between the layers. $\hat{y}^{(i)}$ is the prediction of the model.

species in the composition space of an APT dataset.

### Deep Neural Networks

Deep neural networks (DNN) mimic the structure of neurons in the human brain [51]. A DNN model consists of layers of neurons that are stacked together. The model is made up of three types of layers, namely: input, hidden and output. The input layer takes in the input data, the hidden layer comprises stacks of interconnected neuron layers, and the output layer makes the final predictions. The input to a neuron layer comes from the output of the preceding layer. A DNN model is shown in Figure 2.3; the interconnections between the neurons are indicated by arrows.

To understand the working of a DNN model, it is essential to look into the structure of a neuron (shown in Figure 2.4). The data that is fed into a neuron through an interconnection is multiplied by the corresponding weight. The weight factor determines the importance of an interconnection between two neurons. Further, a weighted sum is calculated for all the interconnections entering a neuron, and bias is added. An activation function is applied to the weighted sum so that the DNN can capture non-linear relationships between the input and the output. Some of the commonly used activation functions are rectified linear unit (ReLu), sigmoid, and hyperbolic tangent [56]–[58]. The choice of the activation function depends on the data ranges of the input and output and the type of the learning problem. The choice of the activation function can significantly affect the performance of a DNN model. The output of the neuron calculated after applying the activation function serves as input to the neurons in the next layer of DNN.

Further, for the given training data the weights and biases in each neuron are optimized by using backpropagation and gradient descent algorithm so that the DNN model can make accurate predictions for previously unseen new data [51], [58]. In each iteration, the training data moves or propagates forward in a DNN, and a prediction is made at the output layer. The predictions are

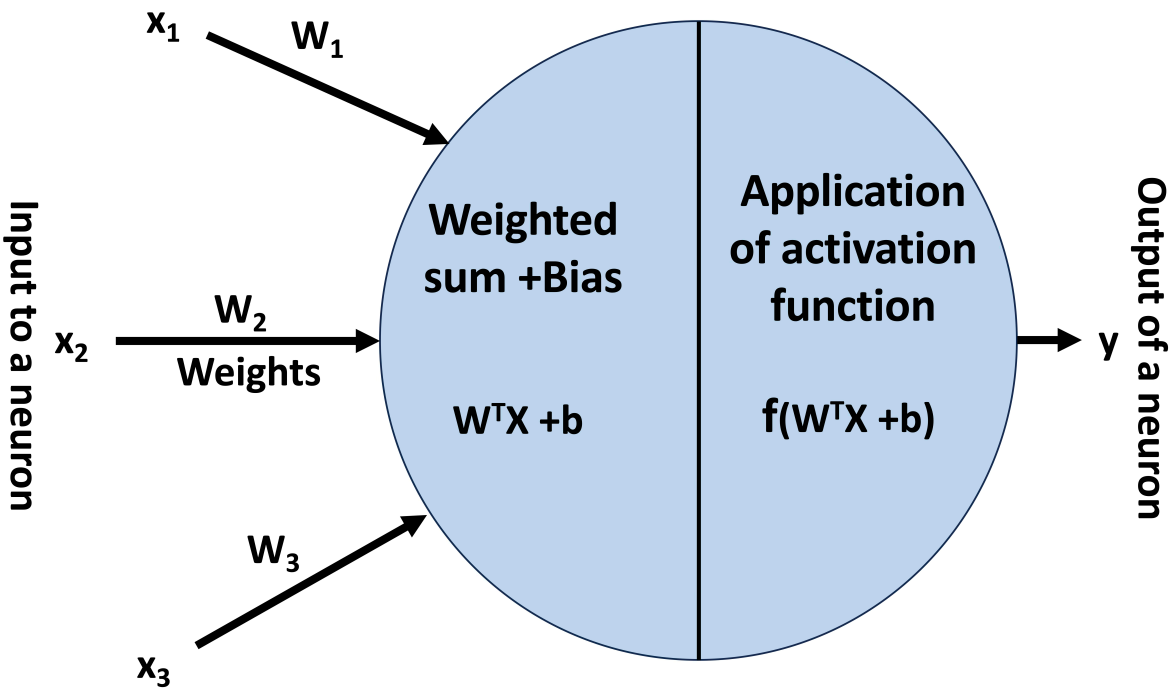



FIGURE 2.4: A schematic depicting the structure of a neuron. $x_1, x_2, x_2$ are the inputs to the neuron, and the $W_1, W_2, W_3$ are the weights associated with the input interconnections (arrows). $X$ and $W$ are the column vectors corresponding to the input values and weights. $f$ is an activation function.

compared to the actual outputs using the cost function, which is calculated for the whole training data:

$$Cost(y^{(i)}, \hat{y}^{(i)}) = \frac{1}{p}\sum_{i=1}^{p}(y^{(i)} - \hat{y}^{(i)})^2 \tag{2.4}$$

where $y^{(i)}$ is the actual output of ith training example. $\hat{y}^{(i)}$ is the predicted output and $p$ is the total number of training examples [51].

The cost function is minimized by tweaking the values of weights and biases using the gradient descent algorithm, while the local gradients at each neuron are calculated by propagating backward in the DNN [58]. DNNs can learn complicated and non-linear relationships between the input and the output data. A DNN model is used in Chapter 6 to find the correlation between the complex SOAP features and the underlying crystallographic information in APT data.

#### U-Net

U-Net is a model that performs classification tasks and is used for semantic segmentation [59]. It was developed for biomedical image segmentation. Since then, it has revolutionized computer vision; for example, it is used in self-driving cars to process optical data. Semantic segmentation means that a class or label is predicted for each pixel in a given input image. Generally, U-Net models perform semantic segmentation by learning the mapping between the input images and human or otherwise annotated masks. Figure 2.5(A) shows a training example corresponding to an electron channeling contrast (ECC) micrograph for a Fe-Mn alloy that was charged with hydrogen, containing dislocation cells and cell boundaries. Figure 2.5(B) shows the annotated image (mask), corresponding to the micrograph in panel (A), where the cells and cell boundaries are delineated using different colors. The training data for a U-Net model consists of such input images and the corresponding

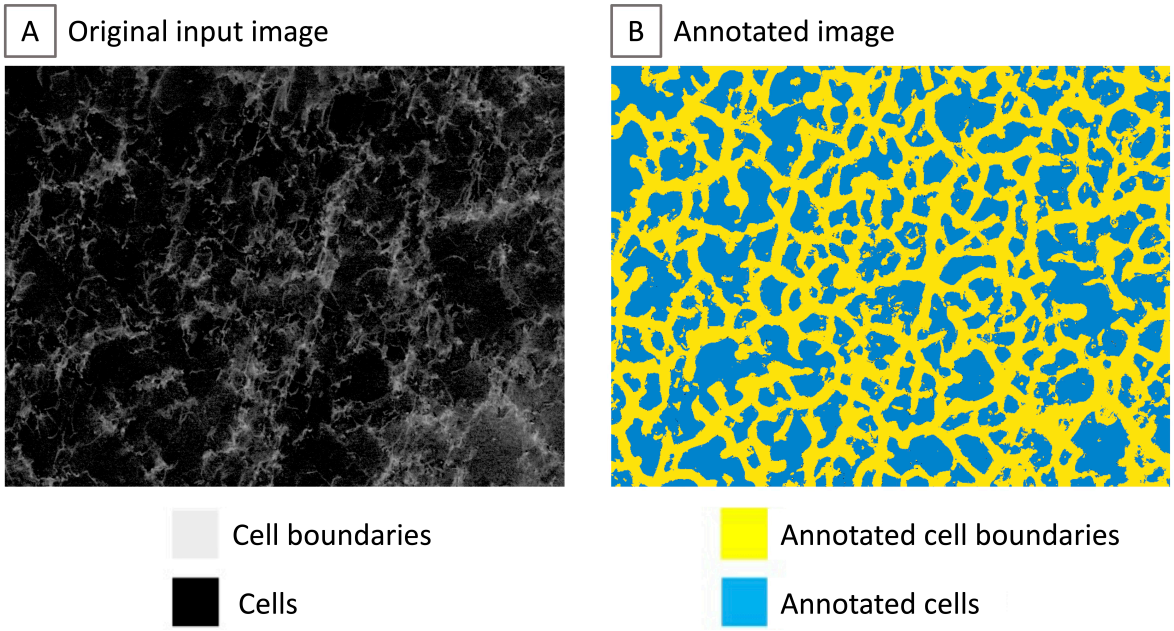



FIGURE 2.5: (A) Electron channeling contrast (ECC) micrograph of a Fe-Mn alloy which is charged with hydrogen. The dislocation cells and cell boundaries are annotated (B) by blue and yellow color. The micrograph and the annotated image can serve as input and output mask images to train a U-Net model.

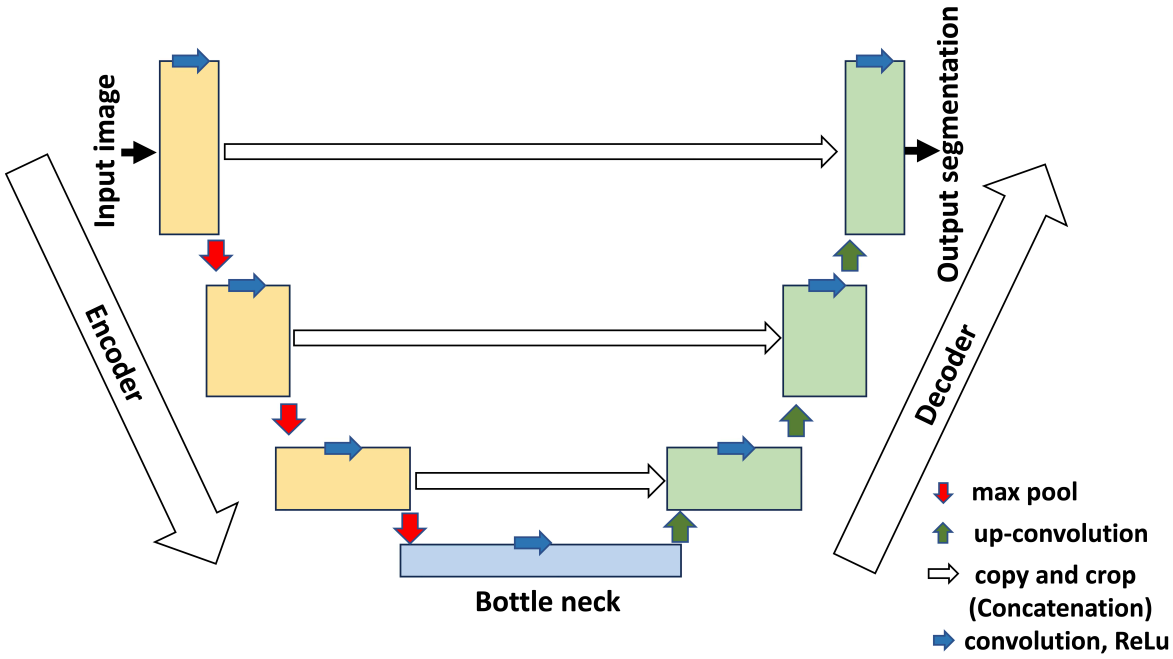



FIGURE 2.6: A schematic of the U-Net model.

ground truth mask images. The learning task for a U-Net is two-fold. First, the model has to learn the features corresponding to a particular object that is classified in the mask image. This is also called contextual information [59]. Second, the model has to learn the features corresponding to the location of the object/class in the mask image. This is also known as localization information [59]. To extract these the U-Net architecture has two parts, namely, encoder and decoder, shown in Figure 2.6. In the encoder part, valid convolution operations, each followed by an application of the ReLu function are applied to the input image in steps, shown by blue arrows in Figure 2.6. Each step is shown by yellow rectangular boxes in the figure. Subsequent to each step a down-sampling is performed to reduce the image size via max-pooling operation, shown by red arrows in the figure. Although the image size is decreased after each step, the number of feature channels is increased by convolution operations, illustrated by an increase in the width of the yellow rectangular boxes after each step. The size of the image is reduced to reach a bottleneck. After this, the decoder part starts in which the size of the image is increased by applying up-sampling via up-convolution operation in steps which are shown by green rectangular boxes in the figure. In each step, convolution operations followed by ReLu are applied. The size of the image is increased to reach the input image size and the number of channels in the output image is equal to the number of the classes in the mask image. The feature channels after each encoder step are concatenated with the corresponding decoder feature channels this is depicted by white arrows in the figure. The encoder part learns the features corresponding to the context information in the mask, for example, for the given micrograph it tries to figure out features that are related to the morphology differences between the cells and cell boundaries. On the other hand, the decoder part tries to learn the location information of the cells and cell boundaries so that semantic segmentation in the image can be performed. Notably, the concatenation step combines the key features that have both the context and location information [59]. The difference between the predicted mask and the ground truth mask is quantified using cross-entropy loss which is minimized using backpropagation and gradient descent algorithms to optimize the U-Net model parameters. The U-Net model is used in Chapter 4 to segment the junction regions from the plate-like regions in 3D APT data.

## 2.2 Unsupervised learning

Unsupervised machine learning algorithms aim to identify patterns within unlabeled input data, enabling them to make decisions or predictions for new, unseen data [53]. These algorithms fall into two categories: clustering and dimensionality reduction, which are elaborated below.

### 2.2.1 Clustering

In clustering, the unlabeled input data is divided into subgroups or clusters each containing similar data points [50]. The similarity can be measured, for example, by calculating the Euclidean distance between the data points. The data points belonging to a cluster will lie near each other. Similarity can also be quantified by checking if the data points are sampled from the same probability distribution. The data points forming a cluster are highly likely sampled from the same distribution. The clustering algorithms used in this thesis are discussed below.

#### Gaussian mixture models

Gaussian mixture models (GMM) are used to cluster unlabeled data using a combination of Gaussian distributions [53], [60]. The number of clusters existing in a dataset cannot be predicted by the GMM

and has to be fed as a hyperparameter by the user. GMM provides the probabilities of each point in the data to belong to a particular cluster or Gaussian. However, for a given set of data points, the representative Gaussian distributions are initially unknown. Each Gaussian is defined by a set of unknown variables (latent variables): $\mu$ (mean), $\epsilon$ (covariance), and $\phi$ (weight of each Gaussian). The task in GMM is to optimize these variables such that the Gaussian distributions fit the given dataset. This optimization problem is solved using the expectation-maximisation (EM) scheme [60], [61]. EM is an iterative optimization scheme used to find latent variables in GMM. If a dataset has $K$ clusters, then there is a need for $K$ Gaussian distributions to represent this dataset. Each Gaussian is initialized with a random set of latent variables ($\mu_k, \epsilon_k, \phi_k$, where $k$ =(1,..., $K$)). The EM algorithm consists of two main steps. The first is the expectation (E) step in which the probability ($P$) that a given data point, $x$, is sampled from a Gaussian distribution with variables $\mu_k, \epsilon_k, \phi_k$ is calculated.

$$P((\mu_k, \epsilon_k, \phi_k)|x) \tag{2.5}$$

This procedure is repeated for each data point and each Gaussian distribution. The second step is called maximization (M). In this step, data points are assigned to each Gaussian cluster based on the probabilities calculated in the previous step. This is followed by adjusting the initial parameters $\mu_k, \epsilon_k, \phi_k$ according to the assigned data points to get new parameters. The E and M steps are repeated until convergence which can be defined as when there is no change in the values of model latent variables. GMM provides soft assignments or probabilities of each data point to belong to a cluster. GMM has been used in Chapter 3 to perform clustering in the composition space to identify different phases/chemical domains in a given APT dataset.

#### Density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN is a density-based clustering algorithm [62]. It extracts points from dense neighborhoods in a given dataset as clusters and the points that correspond to less dense regions are filtered out as noise. Further, DBSCAN model does not require the number of clusters present in a dataset as an input hyperparameter. It can uncover clusters of any shape or extent, making it a robust and flexible clustering algorithm. In DBSCAN, the density information in a region is evaluated using two user-defined hyperparameters: $\epsilon$ and $m$. The local neighborhood around each data point is defined by radius $\epsilon$ [53]. Whereas $m$ designates the minimum number of points required in the local neighborhood defined by $\epsilon$ around a given data point. Three distinct types of data points, shown in Figure 2.7, can be identified based on this hyperparameter characterization, which are listed as follows:

1. **Core point**: It is a point which has at least $m$ number of points around it within a distance of $\epsilon$.
2. **Border points**: It is a point which has less than $m$ number of points around it within a distance of $\epsilon$. However, it lies within $\epsilon$ distance from a core point.
3. **Noise points**: A point that does not belong to the above two categories.

Furthermore, another concept that must be defined for density analysis is reachability. Two data points, $x^{(1)}$ and $x^{(n)}$, are called directly reachable if they are within $\epsilon$ distance of each other. However, they are called reachable if we can define a path $\{x^{(1)}, x^{(2)}....x^{(n)}\}$ such that any two adjacent points in the path, $x^{(j)}$ and $x^{(j+1)}$, are directly reachable.

In the DBSCAN algorithm, a random point, $x$, is initially chosen from the input data. This is followed by a neighborhood analysis of the point $x$. If $x$ is a core point, then a new cluster is started.

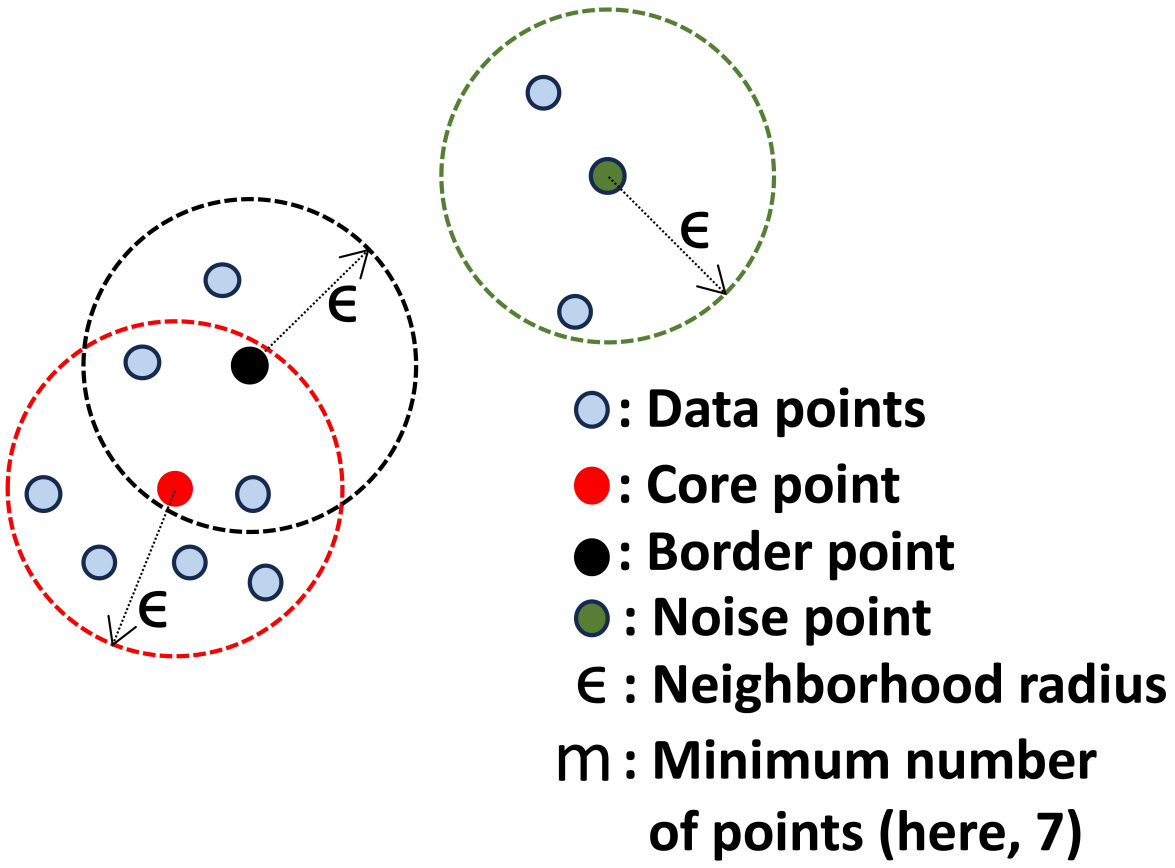



FIGURE 2.7: Core, border, and noise points are shown in an input dataset. $\epsilon$ and $m$ are the radius defining the local neighborhood around a point and the minimum number (here, 7) of points required in the neighborhood, respectively.

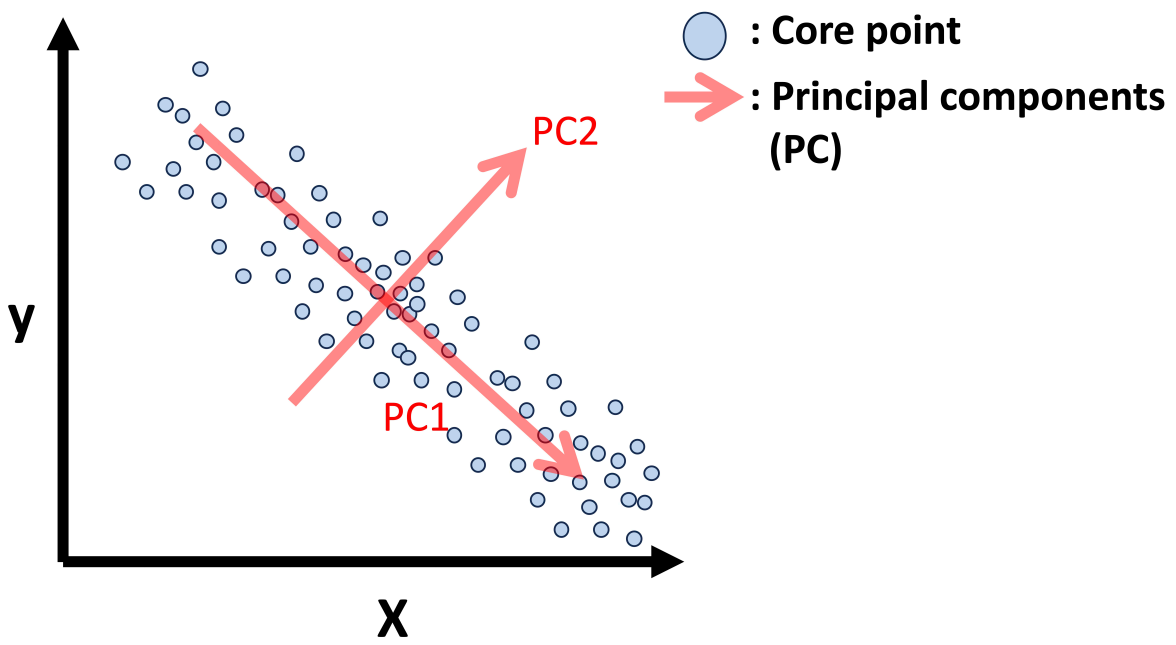



FIGURE 2.8: Principal components of a two dimensional dataset.

On the contrary, if it has less than $m$ number of points within $\epsilon$ distance, it is labeled as noise, and the next data point is visited. This label can change later if $x$ is in close proximity to another core point. After this analysis, the points that are reachable to point $x$ (if core point) are added to the new cluster to expand it. These steps are repeated until all the points in the data are visited. DBSCAN is used in chapters 3 and 4 to cluster various microstructural features in APT data at voxel resolution.

### 2.2.2 Dimensionality reduction

Dimensionality reduction techniques focus on transforming data from a higher-dimensional space to a lower-dimensional one. These methods are essential for effectively visualizing complex data and preventing overfitting in supervised learning models trained on high-dimensional datasets [51], [53]. Various well-known dimensionality reduction algorithms exist, such as t-distributed stochastic neighbor embedding (t-SNE) [63], uniform manifold approximation and projection (UMAP) [64], principal component analysis (PCA), and autoencoders, among others. In this work, particular emphasis is placed on autoencoders (elaborated in detail in Chapter 6) and principal component analysis.

#### Principle component analysis (PCA)

PCA involves employing a linear orthogonal transformation on the given $D$ dimensional input data into lower dimensional space in a way that the variance of the transformed data is maximized [60], [65]. If the lower dimensional space has $Q$ dimensions/features, then each of these dimensions, also called principle components (PC), will be formed by a linear combination of the original dimensions [53]. PC vectors for a two-dimensional input dataset are shown in Figure 2.8. In the PCA algorithm, first data values belonging to each feature in the given input data are standardized. This means that each feature is scaled such that its mean is 0 and standard deviation is 1. This is followed by the calculation of the covariance matrix that encompasses the covariance between each pair of features in the original data. The covariance matrix is then used to calculate the eigenvectors and eigenvalues. Eigenvectors correspond to the directions of maximum variance in the given data, and eigenvalues

depict each direction's weight or importance. The top $Q$ most important eigenvectors are selected as the principal components. A projection matrix is formulated based on principal components and the original data is transformed into a $Q$ dimensional space. PCA is used in chapters 3,4 and 5 to find the longest direction (direction with maximum variance, PC1) in 3D microstructural features. PCA formed the basis for segmenting complex microstructural features based on their geometry.

# Chapter 3

# Chemical domains segmentation: Clustering in composition space-based approach

*This chapter is a modified version of my publication [48]. The Fe-doped Sm-Co alloy APT dataset was provided by Leopoldo Molina-Luna, Oliver Gutfleisch, Nikita Polin, and Baptiste Gault. Aparna Saksena provided the Ni-based superalloy APT dataset.*

## 3.1 Introduction

APT provides insights into the composition of a variety of microstructural features discernible through their chemical (composition) fingerprint, including grain boundaries, dislocations, and secondary phase precipitates [11], [12], [13]. The interplay between these microstructural features along with their chemistry determines the material's macroscopic properties. For exploring structure-property relations and, ultimately, for tailoring new materials, the individual objects forming the microstructure must be quantified in abundance, size, composition, geometrical shape, etc.

Obtaining such a comprehensive quantitative description from APT datasets in a reliable, controlled, and reproducible way remains challenging. The size of the data sets has steadily increased with a wider field-of-view [16] and the implementation of laser pulsing capabilities [17], [18]. APT datasets now routinely contain 10s - 100s of millions of ions. This limits the choice of the applicable algorithms. Quantifying the microstructural features including their composition, volume fraction, size, geometry, or spatial distribution of secondary phases, requires first some form of segmentation of the APT data [20]. The agglomeration of chemical species at different microstructural features within APT data can facilitate this segmentation. This means that atoms of certain species are on average closer together, i.e., the local density of atoms of a specific species can be used to distinguish microstructural features [66]. Certainly, the density may be affected by aberrations in reconstruction due to the differences in phases' evaporation field or chromatic aberrations [67]. Therefore, one should be cautious in interpreting the results.

The most common approach in this direction is the use of iso-composition surfaces, based on a marching cubes algorithm [68], that delineate microstructural features above a threshold composition compared to their surrounding, facilitating visualization. Iso-composition surfaces, often called isosurfaces, provide quantitative information about the composition, and qualitatively define the geometry and distribution of the microstructural features. The caveat is that, as with most aspects of APT data processing, the information from isosurfaces may be sensitive to user-defined parameters.

These parameters are often chosen ad hoc, e.g., to highlight specific characteristics of the data. This may lead to inconsistencies when a single dataset is analyzed by different scientists [69],[70], [71].

More recently, various clustering algorithms, as a complementary approach to isosurfaces, have been proposed as a more controlled alternative, for instance GMM (Gaussian mixture model) [28], DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [27], HDBSCAN (Hierarchical DBSCAN) [72], and OPTICS [73]. Clustering algorithms applied directly to the atomic coordinates of APT datasets tend to be computationally expensive due to the large number of atoms. Therefore, the impact of the clustering algorithm's hyper-parameters (e.g. number of components or clusters in GMM) on the outcome of this process is rarely evaluated systematically.

Here, I propose an approach based on unsupervised machine learning (ML) to semi-automatically extract "chemical domains", i.e. regions with a characteristic composition that corresponds to different phases and segregation zones in the APT data. Contrary to performing clustering on the spatial coordinates of the atoms, I perform clustering in a space containing local composition information of the APT data, with the possibility to also provide the atom clusters in the point cloud within a few minutes on a single CPU core even for datasets containing 500 million atoms. This approach is showcased in the analysis of APT reconstruction data corresponding to a Fe-doped Sm-Co hard magnet, whose magnetic properties depend on its microstructure. The application of the developed workflow helps to identify and segment different phases into separate entities for thorough composition analysis. The developed technique is also qualitatively compared to traditionally employed isosurfaces using a synthetic APT dataset. Finally, the limitations of the technique to isolate various phases in a Ni-based superalloy containing many chemical species are discussed and addressed using a random forest algorithm.

## 3.2 Materials and Methods

### 3.2.1 APT Data

#### Experimental Data

The APT reconstruction data corresponding to a production-grade Fe-doped Sm-Co permanent magnet is used to show the key features of the developed unsupervised learning approach. The production-grade Fe-doped Sm-Co alloy is prepared by first milling and crushing the book mold ingots and subsequently mixing the prepared powders to achieve the desired composition (wt%) $Sm_{25}Zr_3Co_{49}Fe_{19}Cu_5$. This is followed by isostatic pressing and sintering of the powders. For further details see Sample synthesis in [74].

Specimens for APT were prepared by focused ion beam on a Dual-Beam Helios Nanolab 600i System using the approach outlined in [75]. APT measurements were performed on a CAMECA local electrode atom probe (LEAP 5000 XS) in laser pulsing mode, with 10 ps laser pulses at a wavelength of 355 nm (UV), 45 pJ pulse energy at a repetition rate of 200 kHz, in ultra-high vacuum ($1 \times 10^{-10}$ mbar). Further, the specimen base temperature was kept at 60 K with $6 - 10$ ions detected per 100 pulses on average.

Further, reconstruction data for Ni-based supper alloy with the composition given in the table 3.1 used to showcase how the developed methodology behaves for alloys with a large number of elemental and molecular species. The experimental details for APT measurements and specimen preparation are elucidated in [76] .

TABLE 3.1: Composition of Ni-based superalloy in at.%. The rest of the amount is Ni.

| Cr | Co | W | Al | Ti | Re | Ta | Mo | Hf |
|---|---|---|---|---|---|---|---|---|
| 7.50 | 9.90 | 2.20 | 12.7 | 1.3 | 1.0 | 2.2 | 0.4 | 0.03 |

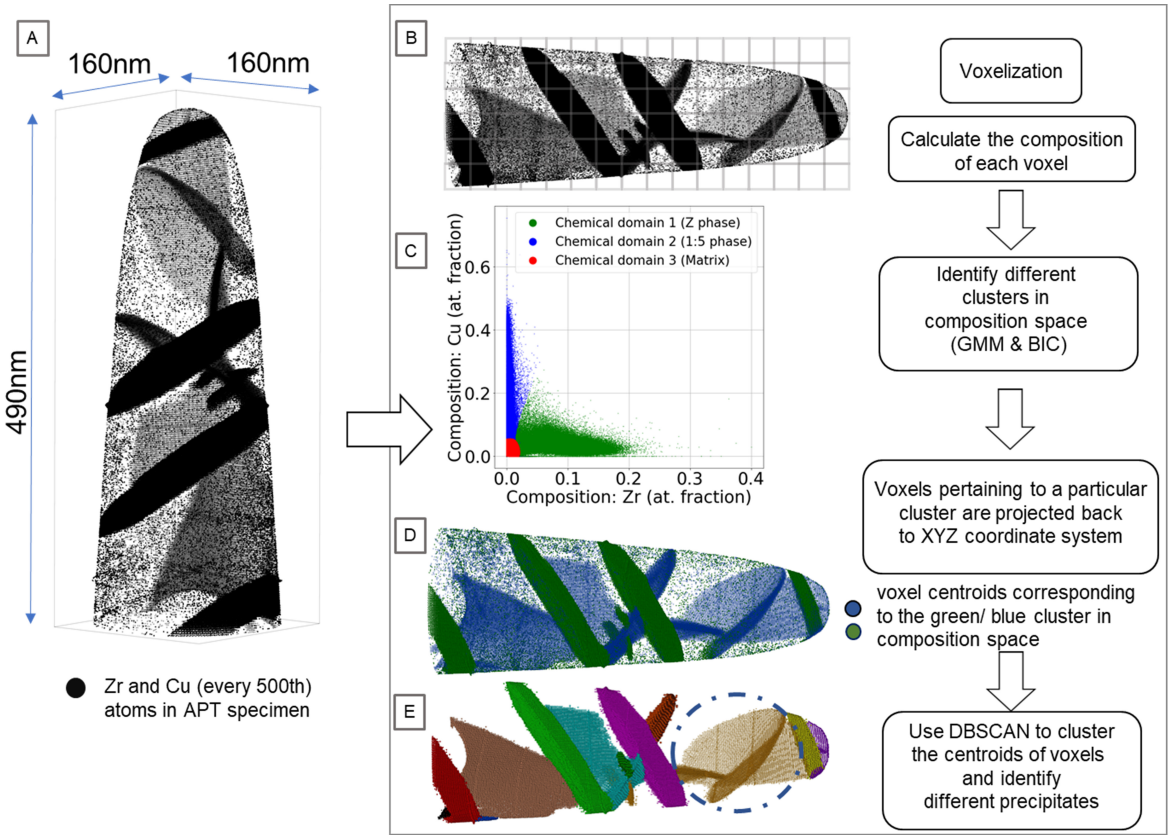



FIGURE 3.1: **(A)** APT reconstruction for a Fe-doped Sm-Co alloy. For clarity, only minority elements (Cu, Zr) are shown. **(B)** Voxelization and per-voxel composition analysis, the grey grid schematically illustrates the voxels. **(C)** 2D projection of the 5D composition space. Each point corresponds to one voxel. Color coding according to 3 composition clusters identified (chemical domain 1 (green), chemical domain 2 (blue), and chemical domain 3 (red) ). **(D)** Phases in real space at voxel resolution according to composition classification. **(E)** DBSCAN is used to cluster voxel centroids belonging to each phase to recognize separate individual precipitates. The circled precipitate has a complex morphology.

#### Synthetic data

To perform a rigorous quantitative comparison between commonly employed isosurfaces and the newly developed method, a synthetic APT data set was generated. First, a crystal supercell (60x60x60), with no added noise, with a spherical $L1_2$ precipitate ($Al_3Li$) embedded in an AlLi solid solution (Al: 90 at.% and Li 10 at.%) is constructed. The radius of the $L1_2$ precipitate is 15nm. Subsequently, noise is added to the positions of the atoms using a Gaussian distribution with a standard deviation of 0.4nm in the x and y directions and 0.2nm in the z direction. Lastly, 20% of atoms are randomly removed to simulate limited detection efficiency in real APT data, see section 3.3.3.

### 3.2.2 Classification of chemical domains: clustering in composition space

To identify the chemical domains present in the Fe-doped Sm-Co alloy APT dataset, a composition-informed segmentation algorithm to visualize quickly and statistically understand clusters or precipitates and segregation zones in the APT data is introduced. The algorithm is summarized in Figure 3.1.

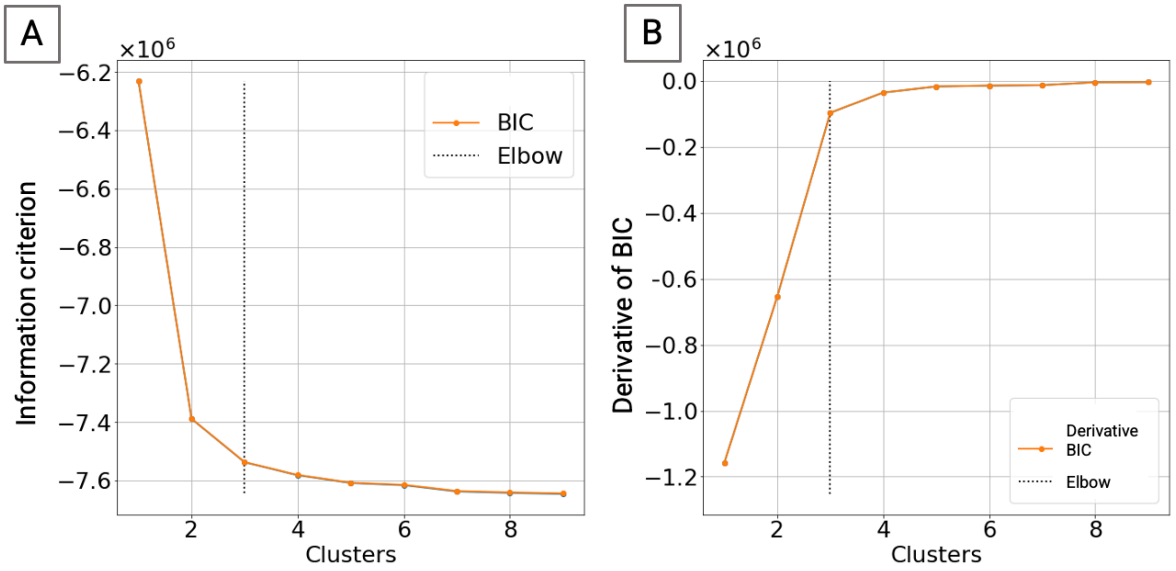



FIGURE 3.2: **(A)** An elbow, shown with a dotted line, at clusters = 3 is observed in the BIC curve. The number of clusters is in agreement with the known phases of the given material.**(B)** The central difference numerical derivative of BIC is invariant from clusters = 4.

The APT data set is divided into equally-sized voxels (here: $2 \times 2 \times 2nm^3$), without application of any delocalisation scheme [20]. For each voxel, the composition, i.e., the relative fraction of all chemical species, is calculated as a vector (see Figure 3.1(B)). These vectors describe points in a multi-dimensional unit simplex which we call composition space. The number of dimensions of the composition space depends on the number of relevant chemical species in the APT dataset. In this showcase, the APT dataset consists of five elements (Sm, Co, Fe, Zr, Cu), so the composition space is five-dimensional. The rationale behind the proposed method is that the voxels that are present inside a precipitate or cluster, of a particular phase or segregation zone, will have approximately the same composition and therefore appear as clusters in composition space. These "clusters in composition space" imply that many voxels exhibit similar composition; they must not be confused with spatial clusters of atoms or ions that form e.g. in segregation zones. The size of voxels is a compromise between spatial resolution (smaller voxels can map smaller features) and reducing statistical noise (larger voxels have less scatter in the compositions and better distinguish chemical domains of close compositions) [77]. While I use cubes of side 2 nm in this case study, this should be adapted for other datasets, depending on the complexity of the underlying phases and length scale of phase separation. For optimization of the voxel size for a given APT dataset, it is recommended to follow the guidelines given in [78].

The second step is to identify different clusters in the composition space. For this, the GMM clustering algorithm [50] is used, see Figure 3.1(C). The rationale behind choosing GMM algorithm is that deviation in composition arises from stochastic effects (e.g. due to intrinsic disorder in the material or the limited detection efficiency) and in limit of large numbers the underlying distribution becomes normal. The GMM algorithm cannot automatically guess the number of clusters in a given dataset. This information has to be fed to the algorithm as a hyperparameter. The approximate number of clusters or chemical domains in the composition space can be determined using the Bayesian information criterion (BIC) or can be inferred from, e.g., thermodynamic databases. In the current work, I have used minimization of the BIC to get the number of clusters as shown in Figure 3.2(A) [53]. For the given APT datasets, we get a visible kink or elbow [28], i.e., the number of clusters after which the decrease in the BIC value is not significant (threshold derivative of BIC = $-0.1 \times 10^6$), at

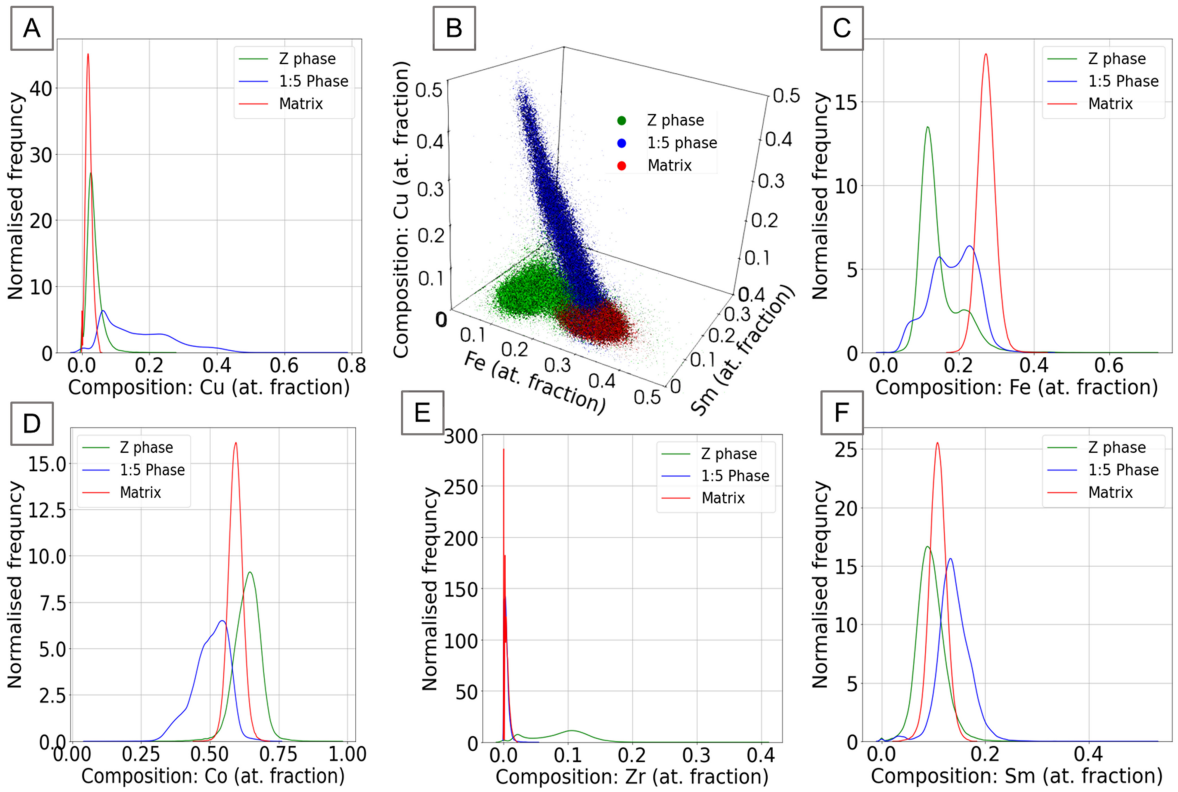



FIGURE 3.3: Normalized frequencies of voxel composition (at. fraction) of **(A)** Cu, **(C)** Fe, **(D)** Co, **(E)** Zr, and **(F)** Sm, respectively, for each phase present in one of the Sm-Co alloy APT dataset. **(B)** 3D projection (Cu-Fe-Sm composition) of the 5D composition space with clusters pertaining to respective phases.

three clusters. They physically correspond to the 2:17 phase (matrix), Z phase (rich in Zr) and 1:5 phase (rich in Cu). The number of clusters is also in agreement with the known phases of this material [74]. The kink or change in slope corresponding to the derivative threshold = $-0.1 \times 10^6$ can be seen in the central difference numerical derivative of the BIC as shown in Figure 3.2(B). Figure 3.3(B) shows a 3-D (Cu, Fe, Sm) slice of the 5-D composition space where each point corresponds to a voxel in the spatial coordinate system. The three clusters or chemical domains found by the GMM algorithm are marked in Figure 3.3(B). Further, normalized frequencies of voxel composition for each chemical species shown in Figures 3.3(A, C, D, E, F), obtained using kernel density estimate (kernel-bandwidth $= 0.1$), highlight that no single element is sufficient to assign a voxel to a certain phase, but definitely their combination is required. The voxels pertaining to a particular cluster in composition space are translated back to the spatial coordinate system so that voxels present in the matrix, precipitates, and segregation zones can be separated, see Figure 3.1(D).

After 3D phase segmentation using clustering in composition space, DBSCAN is applied separately to the centroids of the voxels belonging to the same composition space cluster [62]. Here, a voxel centroid is the arithmetic mean of the positions of the atoms present in the given voxel. In this process, the noisy voxels, which often lie at the boundary of the dataset, are removed and various precipitates pertaining to each phase are identified as shown in Figures 3.1(D and E). Since both the atoms residing in each voxel and the voxels present in each precipitate are registered, the last step enables us to study each precipitate separately. Moving from an atom-by-atom representation of the data set to voxel centers is particularly advantageous for DBSCAN since the atomic density is subjected to fluctuations due to the APT experiment (e.g., pole formation [79] or inhomogeneous evaporation due to evaporation-field differences [80]), while the voxel density is uniform. The voxel

TABLE 3.2: Volume fraction, average thickness and average composition of 2:17, Z, 1:5 phase in one of the given APT datasets.

| | | | Composition at% (mean $\pm$ 1.96 standard error) | | | | |
|---|---|---|---|---|---|---|---|
| Phase | Volume fraction | Thickness(nm) | Sm | Co | Fe | Cu | Zr |
| 2:17 | 0.8 | - | $10.9 \pm 0.4$ | $59.4 \pm 1.0$ | $27.2 \pm 0.8$ | $2.0 \pm 0.8$ | $0.5 \pm 0.2$ |
| 1:5 | 0.06 | $17.6 \pm 4.0$ | $14.4 \pm 1.6$ | $48.0 \pm 5.3$ | $16.1 \pm 5.1$ | $21.1 \pm 8.8$ | $0.4 \pm 0.2$ |
| Z | 0.14 | $10.9 \pm 1.4$ | $9.7 \pm 1.4$ | $63.4 \pm 2.5$ | $14.5 \pm 1.4$ | $3.8 \pm 1.4$ | $8.5 \pm 1.74$ |

centers, by construction, are equally spaced and provide a practically constant density of points within each precipitate. Since now the number of voxels that are present in each precipitate is known, statistical descriptors like mean volume, number density, and nearest neighbor distribution of the precipitates can be readily obtained. For the present analysis, I have used scikit-learn's implementation of DBSCAN (parameters eps or (maximum distance between voxel centroids for one to be considered in the neighborhood of the other) = 9 nm and min_samples (minimum number of voxel centroids required around a particular centroid to identify it as a core point) = 80 are used for all the identified phases) [81].

## 3.3 Results and Discussion

### 3.3.1 Visual validation with isosurface results:

The proposed approach of clustering in composition space has been applied to the given Fe-doped Sm-Co alloy APT dataset. Figure 3.4 provides a visual comparison between the isosurfaces obtained using AP suite and the results of our new algorithm. Figure 3.4(A) shows the 6 at% Zr (Z phase) and 12 at% Cu (1:5 phase) isosurfaces, and Figure 3.4(B) shows the voxel centroids that lie in Z phase and 1:5 phase. These phases have been identified by the composition space clustering, while the final assignment of the voxels as part of a spatial cluster additionally relies on the DBSCAN algorithm on the voxel centroids of each phase to eliminate noise. The phases obtained in Figures 3.4(B) are visually identical to the isosurfaces for the respective phases shown in Figure 3.4(A)

For clearer visualization of the precipitates, they are separated in subfigures (C) and (D): Figure 3.4(C) depicts only the 1:5 phase precipitates, and Figure 3.4(D) the ones for the Z phase. The cells formed by the matrix phase can also be isolated by applying DBSCAN on the voxel-centroids of the 2:17 phase, shown in 3.4(E). The precipitates of 1:5 phase with complex morphologies containing bi and quad junctions are reduced into simpler plate-like morphology by eliminating the junction voxel-centroids which is discussed in detail in the next chapter (Section: **Decomposing geometrically complex precipitates**). A quantitative comparison between isosurfaces and the developed composition space approach is discussed in section **Quantitative comparison with isosurfaces**.

### 3.3.2 Calculation of volume fraction and composition of each phase:

The segmentation of each phase or chemical domain allows for the calculation of the approximate volume fraction, as the number of voxels included in a phase is known. The volume fraction of each phase in the Fe-doped Sm-Co APT dataset is summarised in Table 3.2. The average composition of a phase is the mean of the compositions of the precipitates or domains it comprises. Average compositions and standard error (confidence interval = 95%) for each phase in the APT sample are summarised in Table 3.2. They are calculated by averaging over the subsets encompassing 15000 adjacent voxels each. The dependence of the standard error on subset size for the Z phase is shown

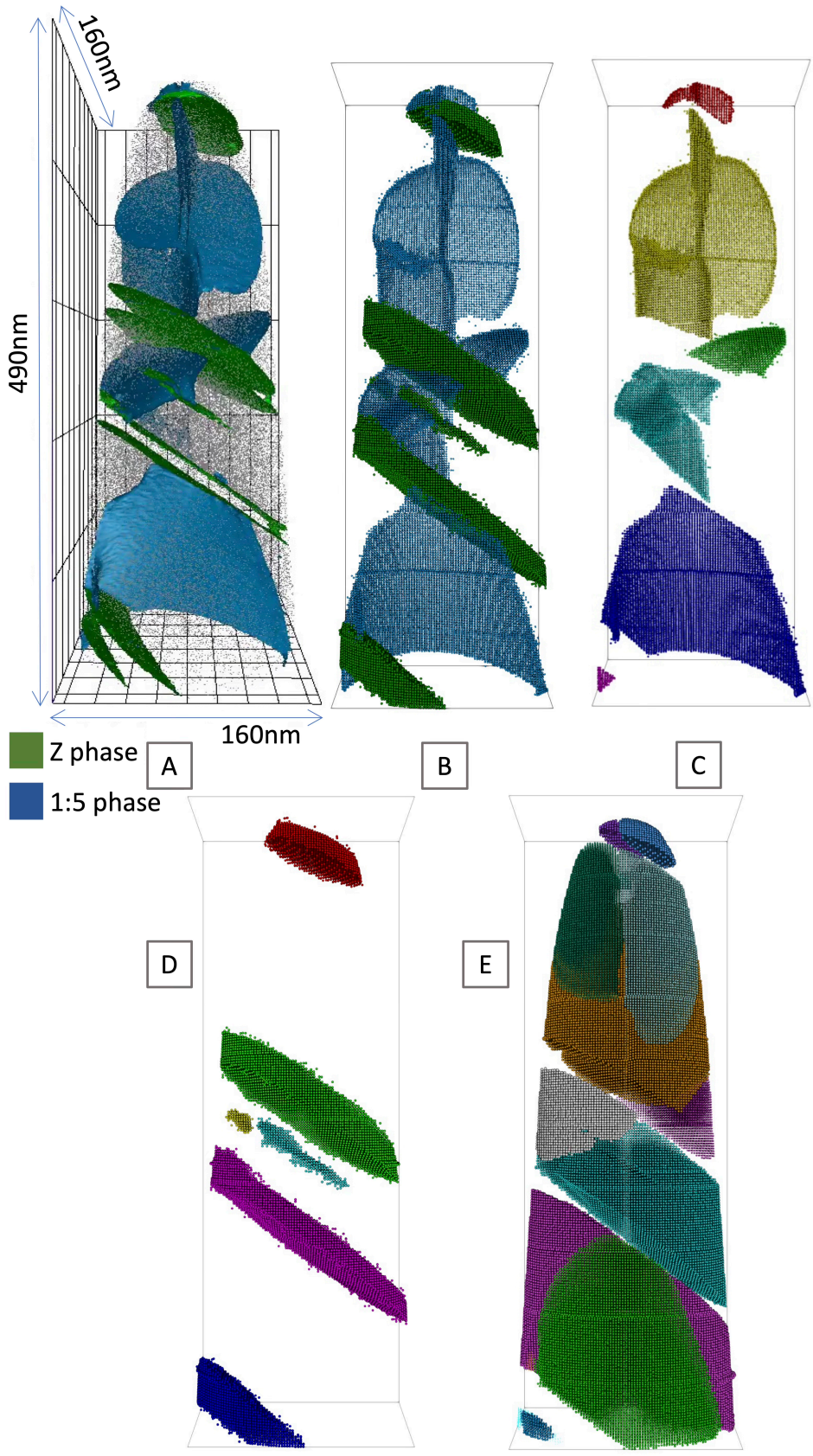



FIGURE 3.4: **(A)** 6 at% Zr (green: Z phase) and 12 at% Cu (blue: 1:5 phase) isosurfaces in a Fe-doped Sm-Co alloy APT dataset extracted using AP suite. **(B)** The voxel centroids lying in the Z phase (green) and 1:5 phase (blue) precipitates extracted using clustering in composition space. **(C) and (D)** Segmented 1:5 and Z phase precipitates respectively . **(E)** Segmented matrix cells after removing Z and 1:5 phase.

in Figure 3.5. As seen in the figure, the standard error associated with a given chemical species demonstrates an inverse relationship with the number of voxels per subset. The standard error then reaches an asymptotic plateau beyond a certain voxel count. It will become zero when the number of voxels in the subset reaches the total number of voxels in a phase.

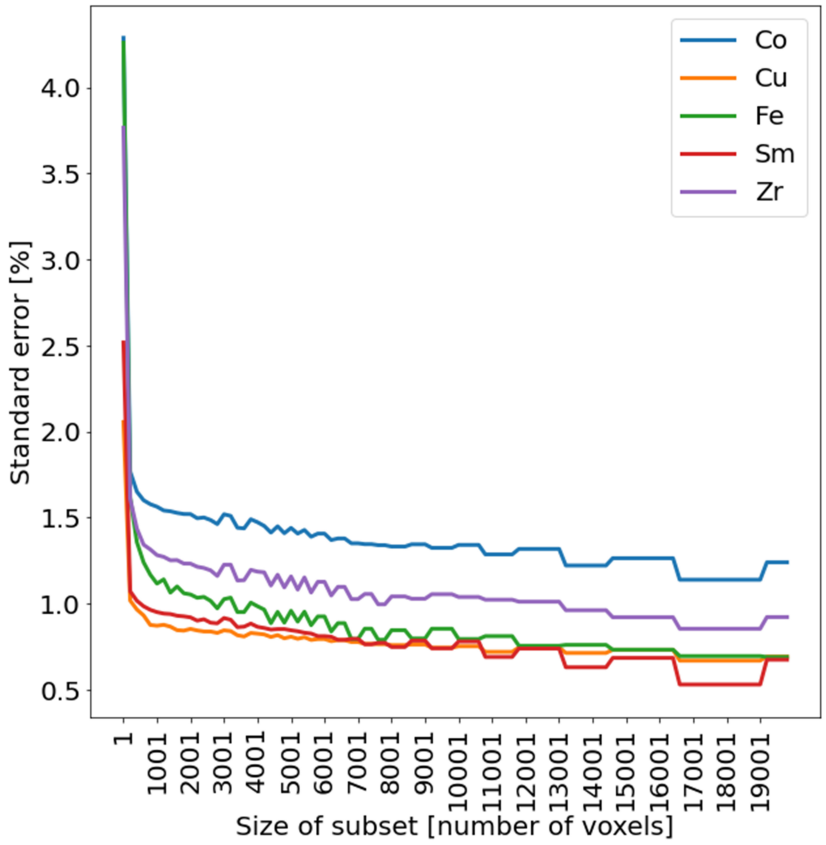



FIGURE 3.5: The dependence of the standard error on subset size (number of voxels) for Z phase in the given Fe-doped Sm-Co alloy.

**Composition fluctuations in precipitates**:

The comparison between the voxel composition distribution of each chemical species in a particular precipitate and the corresponding random distribution are plotted to understand if any clusters or segregation zones are present inside the chosen precipitate. Such a comparison for randomly selected precipitates of the 1:5 and Z phase and a cell of 2:17 matrix phase for Cu, Fe, and Zr is shown in Figure 3.6. As seen in the figure, the experimental distribution deviates from the random distribution for the Z and 1:5 phases while the 2:17 phase shows no such discrepancies and thus appears to exhibit the ideal disorder. In order to understand and quantify these deviations, all the precipitates are reduced to simpler plate-like structures, enabling us to study in-plane composition fluctuations, which is discussed in detail in the next chapter. The computation of volume fraction, the composition of each phase, and composition fluctuations in precipitates can be executed in just a few seconds using a single CPU core for the given APT dataset.

### 3.3.3 Quantitative comparison with isosurfaces

Here, a simple quantitative comparison between the developed composition space approach and isosurfaces has been made. Since the precipitates of the Z phase and the 1:5 phase in the Fe-doped Sm-Co alloy touch the edge of the APT dataset, the corresponding isosurfaces are open also shown

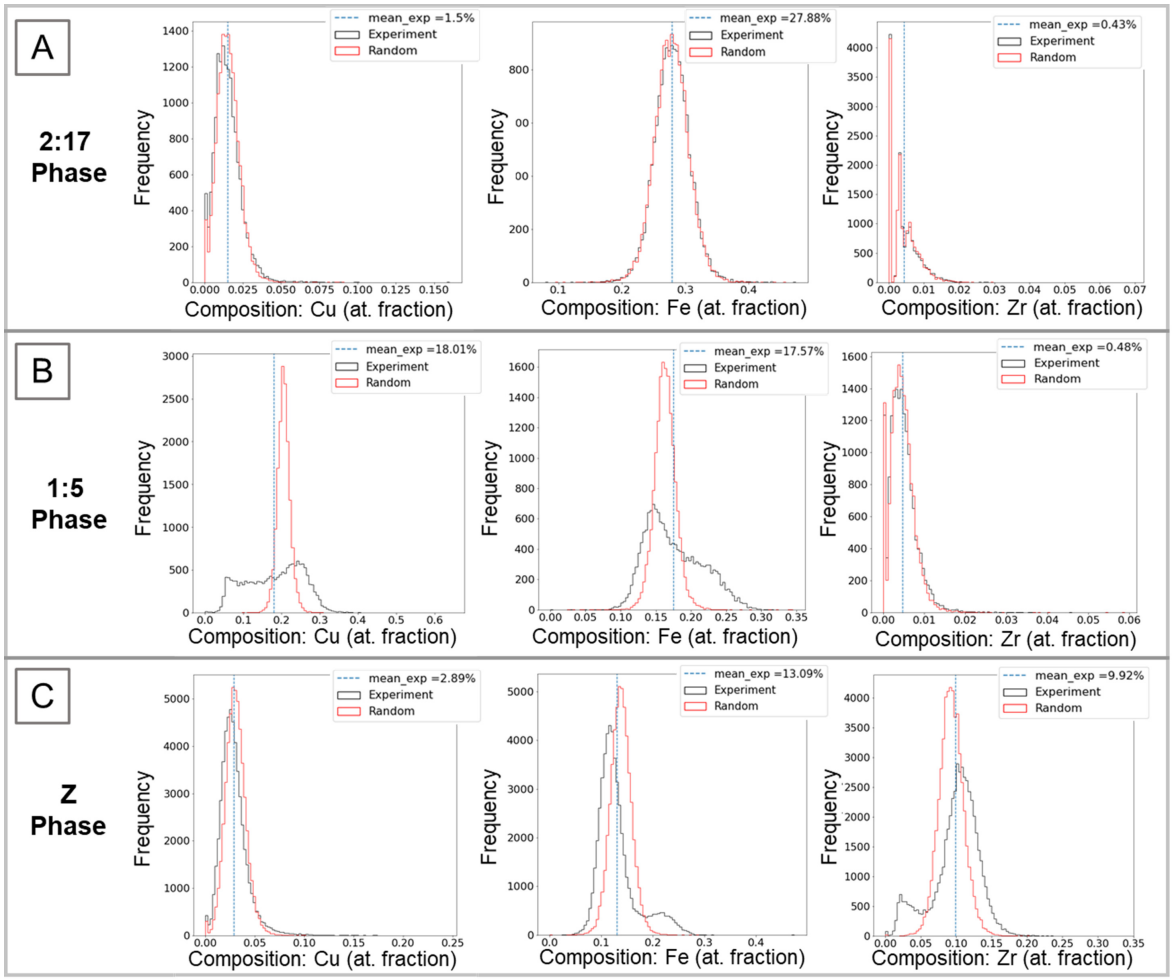



FIGURE 3.6: Voxel composition distributions of Cu, Fe and Zr present in **(A)** a 2:17 matrix cell as well as in 2 precipitates chosen for **(B)** 1:5 phase and **(C)** Z phase. The experimental composition distribution (black) deviates significantly from the random distribution (red) for 1:5 and Z precipitates.

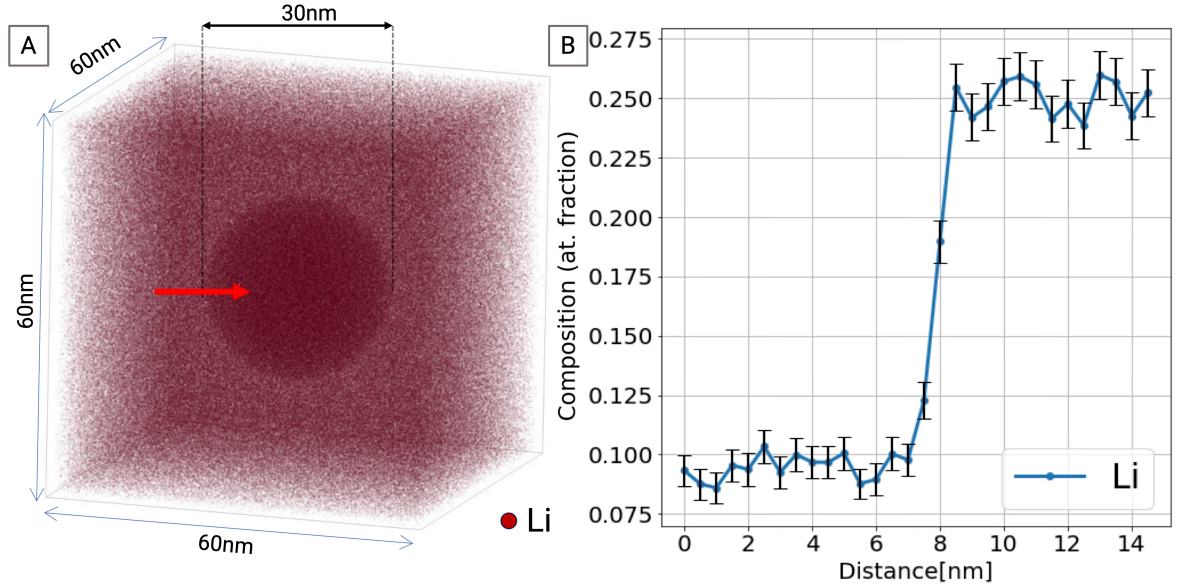



FIGURE 3.7: APT synthetic data consisting of $L1_2$ precipitate ($Al_3Li$) embedded in an AlLi solid solution (Al: 90 at.% and Li: 10 at.%) is shown on left. Only Li atoms are shown. The right panel shows the 1D composition profile (bin size = 0.6 nm) for Li along the red arrow shown in the left panel. The black error bars in the 1D composition profile correspond to the $2\sigma$ counting error.

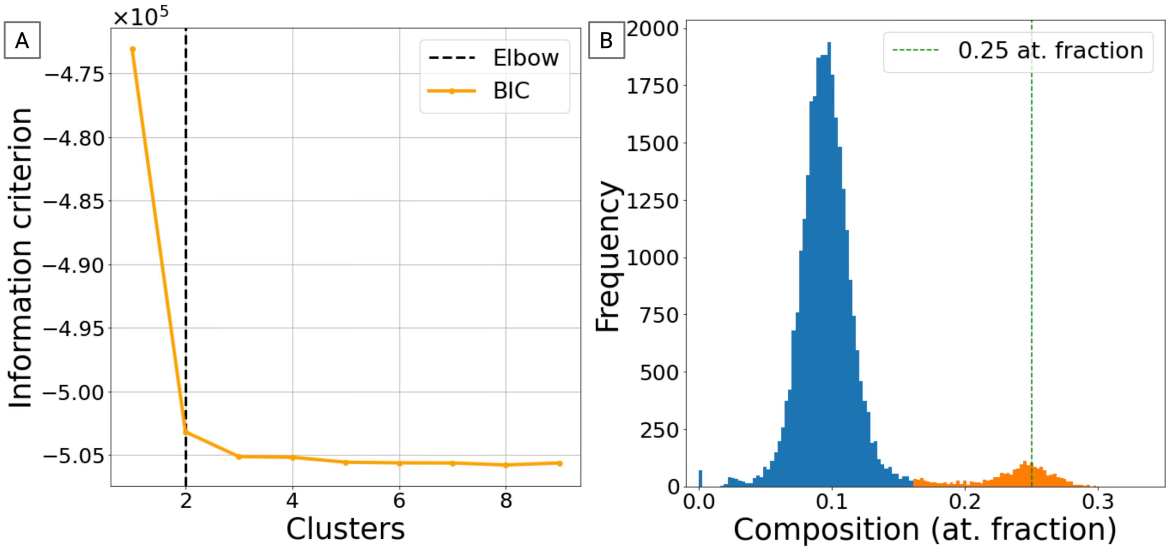



FIGURE 3.8: (A) BIC minimization is applied on the composition space of Al-Li synthetic APT data. The elbow at 2 number of clusters shows that there are two chemical domains or phases in the synthetic data. (B) GMM is applied to the composition space to get the voxels of each phase. The distribution of Li composition in the voxels is shown. The voxels that are highlighted in orange belong to the $L1_2$ phase.

in Figure 3.4(A). Due to these open isosurfaces, any meaningful quantitative comparison with the composition space results, like the calculated volume of the precipitates, is challenging. Therefore, the synthetic APT dataset mentioned in section 3.2.1 is used for the analysis. The dataset comprises a spherical precipitate of $L1_2$ structure ($Al_3Li$) embedded in AlLi solid solution as shown in Figure 3.7(A). In Figure 3.7(B), a one-dimensional compositional profile across the matrix-precipitate interface is presented. This profile shows that the precipitate possesses a lithium atomic fraction (at. fraction) of 0.25, in contrast to the matrix, which exhibits a lithium atomic fraction of 0.10.

The composition-space approach (voxel size = 2nm, DBSCAN [eps = 9 and min_smaples = 80 ] [62]) is applied on the synthetic data. It can be inferred from the BIC minimization plot shown in Figure 3.8(A) that there are two clusters in the composition space (matrix and the $L1_2$ phase). The voxels pertaining to the two identified chemical domains are automatically extracted by applying GMM on the composition space. The distribution of Li composition in the voxels that are segmented into two phases is shown in 3.8(B). The volume of the spherical precipitate is calculated. In addition, Paraprobe [82], an open-source tool for APT data analysis, is used to calculate the isosurafaces (voxel size = [1,1,1] nm, delocalization kernel = [1,1,1] nm) for isovalues from 0.11 to 0.26 atom fraction Li.

Figure 3.9 summarises the volume of various Li isosurfaces and compares them with the geometric (Sphere vol. $= \frac{4\pi r^3}{3}$) and composition space-based volume. Only the maximum isosurface volume at a given isovalue is plotted. As shown in the figure, composition space-based approach slightly overestimates the volume of the spherical precipitate by 2%. This effect may also be attributed to the chosen voxel size. Also, the plot shows the dependence of precipitate volume on isovalue. A change of 0.01 at. fraction in isovalue changes the volume by approximately 3.5% in 0.13 to 0.23 at. fraction range. So, the composition space-based approach enables the segmentation of the 3d microstructure without depending on a sensitive hyperparameter like isovalue as in the case of isosurface calculation.

### 3.3.4 Dimensionality reduction for composition space

As humans live in a three-dimensional world, our imagination or intuitions about high-dimensional data are often not very accurate. Even for machine learning algorithms, the task of finding correlations

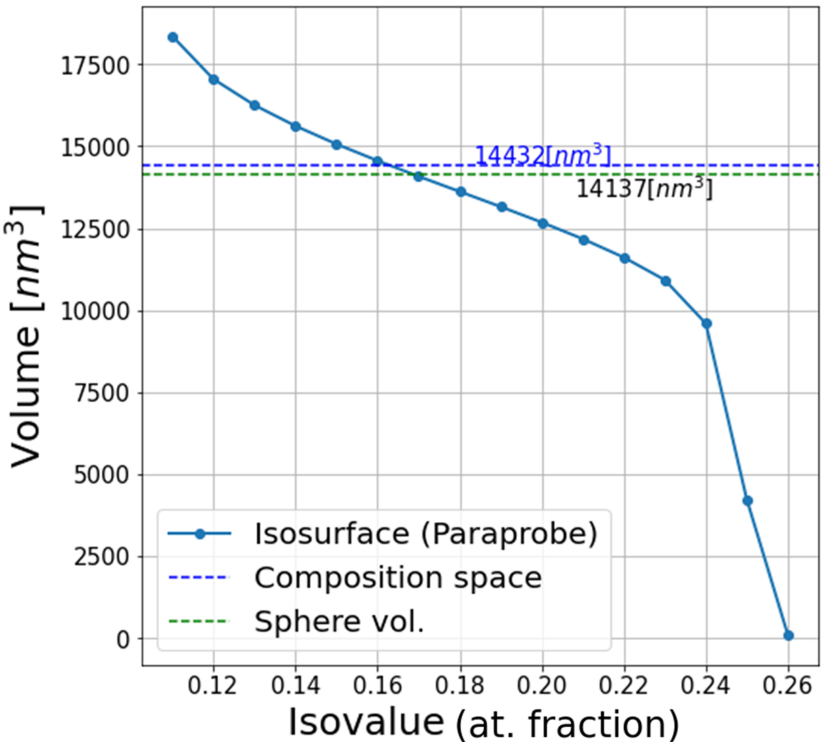



FIGURE 3.9: Dependence of isosurface volume on the isovalue (in the range 0.11 to 0.26 at. fraction Li) for the synthetic data in Figure 3.7. Comparison between the geometric (Sphere vol.) volume and volume of the precipitate calculated using composition space is also shown

in such data is daunting. For example, clustering algorithms, in general, are based on the idea that if certain data points lie in close proximity to each other compared to others, then they form a cluster. In higher dimensional space, however, the Euclidean distance between the same points may be very large. This is also called the "curse of dimensionality" [53], [83]. The GMM algorithm with BIC minimization is used for clustering in the N-dimensional composition space. It is highly probable that as the N increases, the number of clusters predicted by the BIC minimization might not be correct. Consequently, the clustering performed by GMM is not optimum. To make the composition space-based approach robust against this issue, a scheme for dimensionality reduction is developed and discussed below.

As given in Table 3.1, the Ni-based superalloy mentioned in section 3.2.1 has nine elements. Including all the minority elements and the molecular species (oxides, hydrides, and carbides), the composition space of the reconstruction data for this alloy has 31 dimensions or features. This alloy has 2 phases namely: $\gamma$ and $\gamma'$ which are shown in Figure 3.10(A). The $\gamma$ phase is highlighted using 8 at.% Cr isosurface. The BIC minimization plot, which was calculated by applying the GMM algorithm on the composition space for different numbers of clusters, is shown in 3.10(B). As depicted in this figure, the location of the elbow is not apparent. There can be 3, 4, or 5 clusters in the composition space. However, it is already known that there are 2 phases in the reconstruction data. Therefore, the algorithm applied naively on all available species fails to capture the accurate number of phases or chemical domains in the data. This behavior can be attributed to the large number of dimensions (N) of the composition space.

The BIC minimization gives the tentative number of clusters ($C_1$, $C_2$ ... $C_k$) in the composition space, refer to Figure 3.11. For a given number of clusters, $C_k$, the GMM algorithm is applied to the composition space ([V$\times$N], where V is the total number of voxels), and the label ($Y_{pred}$) of each

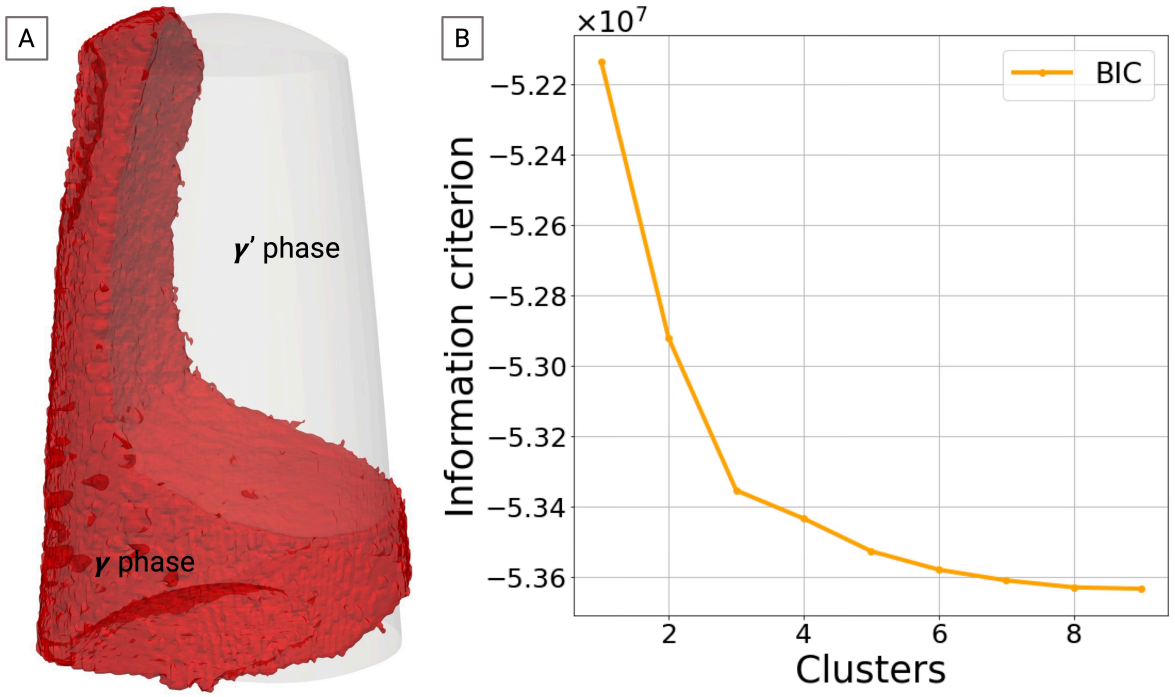



FIGURE 3.10: (A) 8at.% Cr isosurface in the given Ni-based superalloy corresponding to $\gamma$ phase. Rest is the $\gamma'$ phase. (B) BIC minimization plot for the composition space for APT data of the alloy.

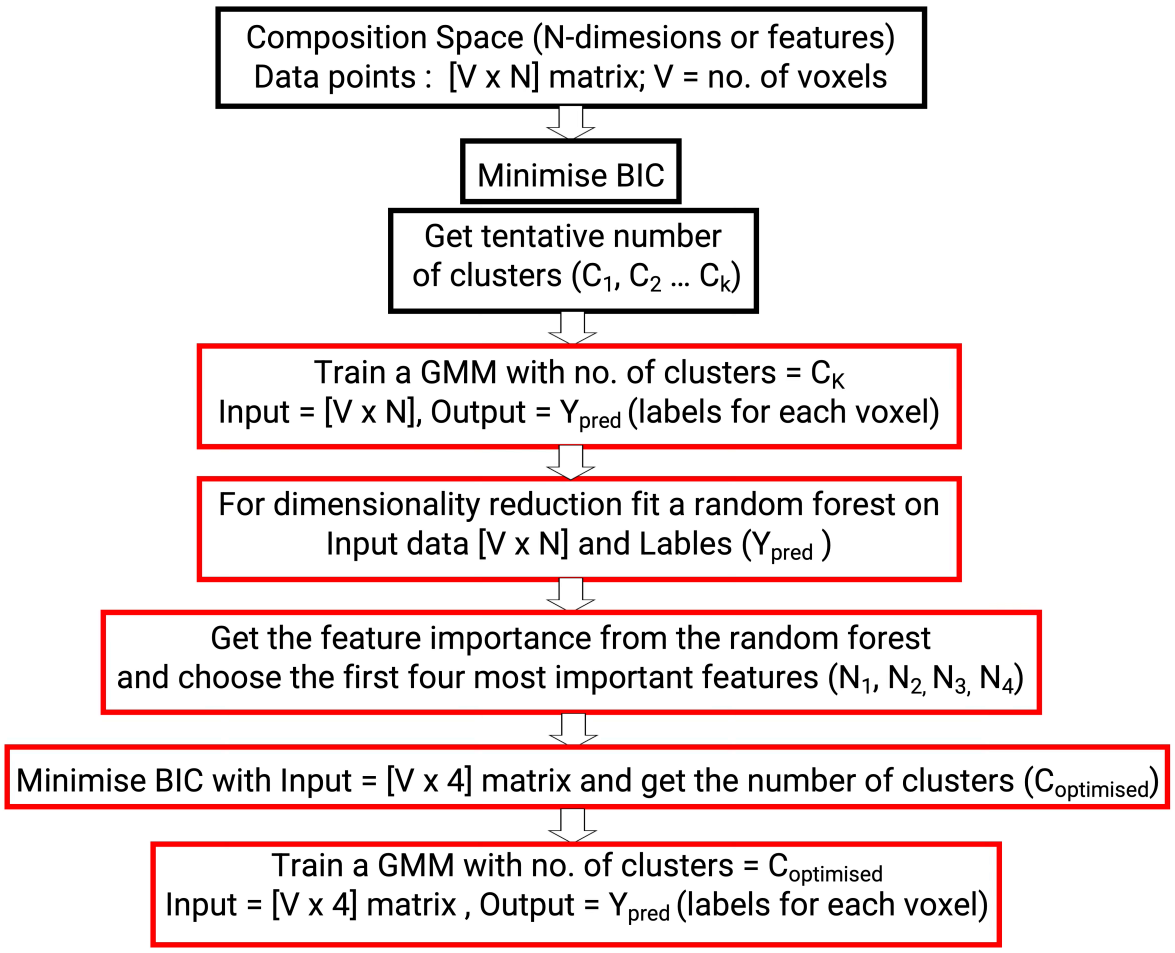



FIGURE 3.11: The developed scheme for composition space dimensionality reduction.

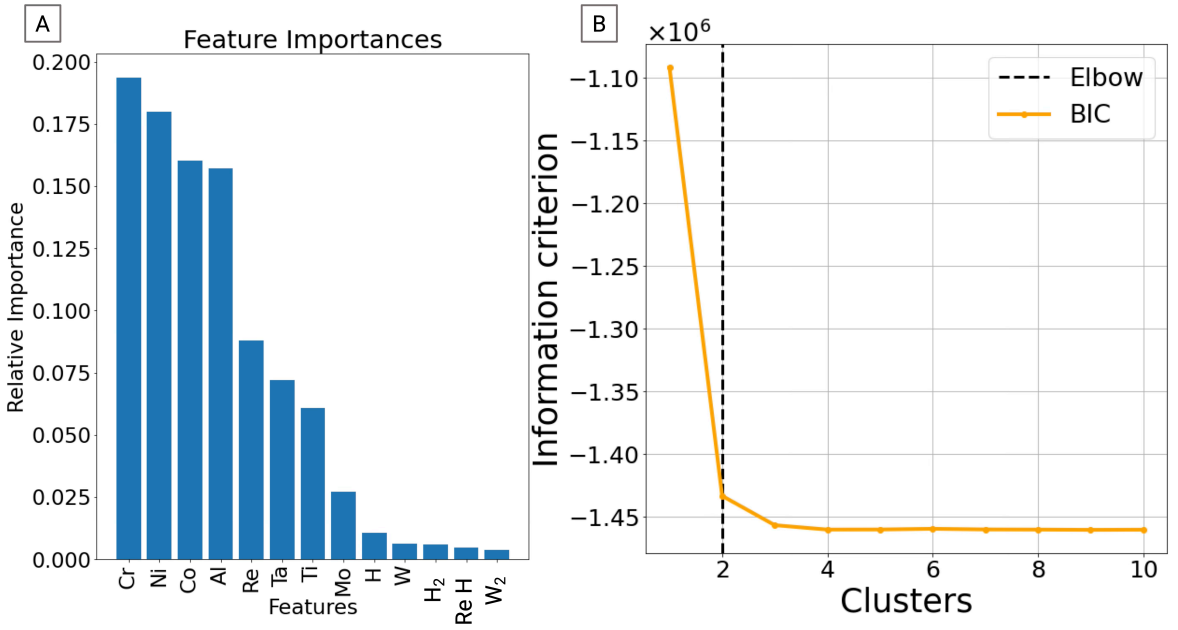



FIGURE 3.12: (A) Relative importance of various chemical species in the Ni-based superalloy considering 3 clusters in the composition space. (B) BIC minimization considering the top 4 most important species/features.

voxel is recorded. Thereafter, a random forest algorithm is used to find correlations between the input composition space ([V×N]) and the predicted labels ($Y_{pred}$) from GMM. The fitted random forest provides the relative importance of features/dimensions in the composition space. Out of N, the top 4 most important dimensions are now used for BIC minimization using GMM to get the optimized number of clusters, $C_{optimized}$. Clustering is again performed using GMM in the composition space ([V×4]), which now has just four features, using the optimum number of clusters. The procedure is repeated for each tentative number of clusters. This scheme is better than using PCA because the random forests algorithm can capture non-linearity in the given data. Secondly, the PCA dimensions are not necessarily the same as the composition space dimensions, which correspond to a chemical species. On the other hand, for better interpretation, random forest directly gives the relative importance of the existing dimensions. This same problem of interpretability exists while applying other dimensionality reduction techniques like t-SNE, UMAP, or autoencoders [53].

The above scheme is applied to the composition space ([V × 31]) of the given Ni-based superalloy. The composition space of the given Ni-based superalloy is 31 dimensional. It can have 3, 4, or 5 clusters according to the BIC minimization plot shown in Figure 3.10(B). To exemplify the working of the scheme, it is initially assumed that the composition space has three clusters. On applying the above scheme to the composition space ([V × 31]), the relative importance of the 31 features is extracted. The top 13 most important dimensions are shown in Figure 3.12(A). After selecting the top 4 most important features, BIC minimization is applied to the composition space([V × 4]). As shown in 3.12(B), there is an elbow at two clusters, clearly suggesting the existence of two phases in the given dataset. The GMM algorithm is applied with 2 clusters to segment the 4-dimensional composition space. A slice of composition space with the two most important features, the composition of Cr and Ni, is shown in Figure 3.13 (A). The centroids of voxels of each phase in 3D after segmentation are shown in Figure 3.13 (D). The same results are obtained upon initially choosing 4 or 5 clusters in the composition space.

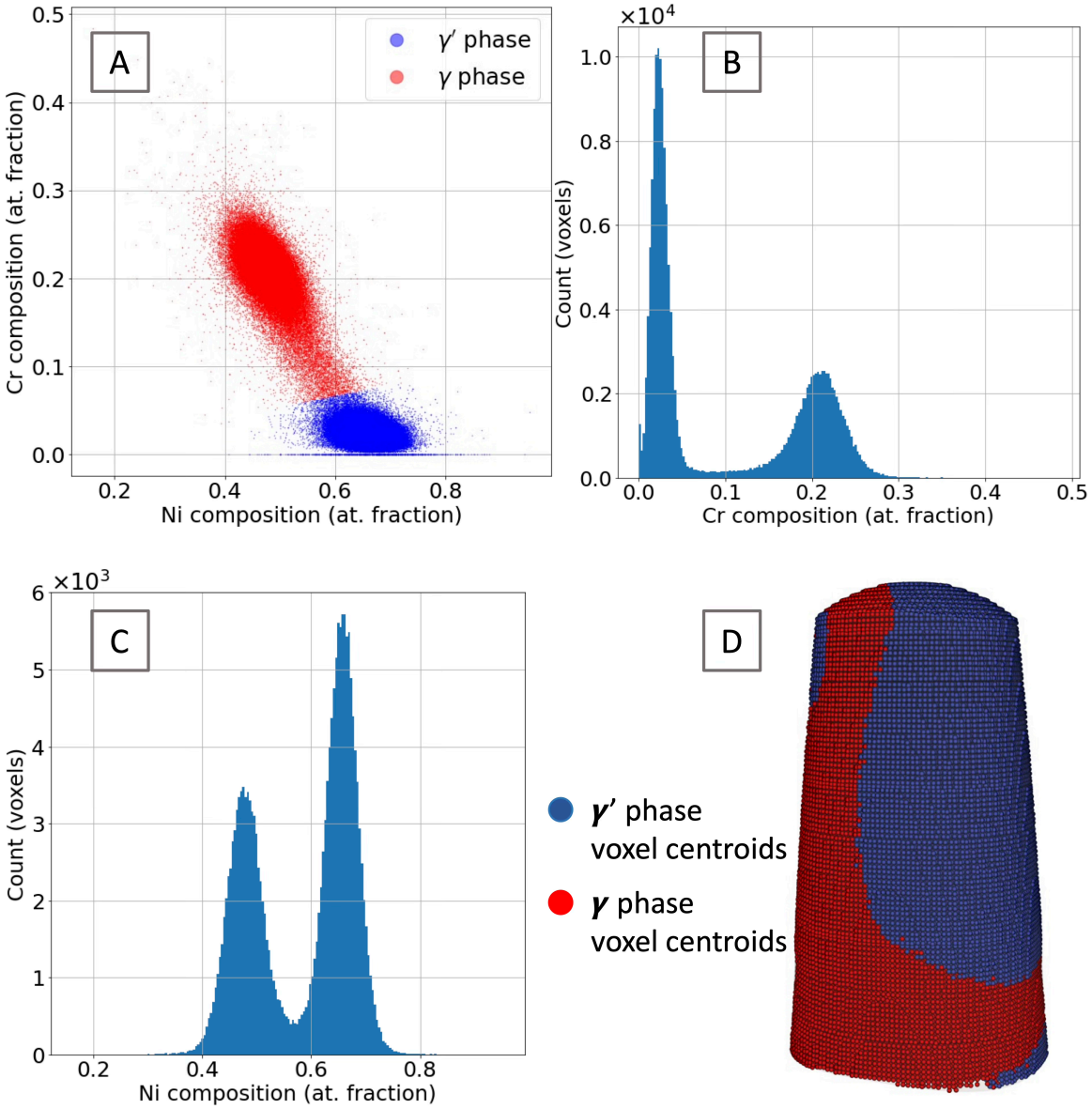



FIGURE 3.13: (A) Clusters ($\gamma$ and $\gamma'$ phase) in composition space considering 4 most important features or species. (B, C) Distribution of Cr and Ni composition in the voxels. (D) Voxel centroids of each identified cluster in composition space in xyz coordinate system.

## 3.4 Conclusion

In this chapter, a Python-based ML workflow to automatically segment out chemically distinct regions from the APT data was developed. The key steps are (1) voxelisation of the APT data set, (2) calculating the composition of each voxel, (3) clustering in composition space to obtain the chemical domains in the dataset and (4) separation of voxels of the extended regions of each chemical domain by DBSCAN clustering algorithm. The relevant data of all steps is stored in hdf5 files to aid subsequent postprocessing. This enables a robust and systematic analysis of the microstructure from its chemical fingerprint, yielding for instance the average volume fraction, domain size distribution, and average composition of the identified chemical domains (which could be stable phases from the thermochemical phase diagram, or segregation zones with a unique composition near grain boundaries or other crystallographic defects). The developed workflow can readily be applied to a wide variety of microstructures analyzed by APT.

A quantitative comparison between the developed composition space-based approach and widely used isosurfaces was also drawn using synthetic APT data. As isosurfaces strongly depend on the isovalues so does their enclosed volume. On the contrary, my present approach depends on clustering in a multidimensional composition space which given enough contrast in composition ensures a more reliable volume for chemical domains. For the synthetic dataset, not only did the composition space-based approach identify the number of chemical domains/phases accurately, but it also computed secondary phase precipitate volume very close (2% error) to the theoretical volume.

Using the developed workflow, three phases, matrix (2:17 phase), Z phase and 1:5 phase, for Fe-doped Sm-Co alloy APT dataset were identified. Notably, this computational analysis can be executed in just a few minutes using a single CPU core, even for APT datasets with 500 million atoms. The complex microstructure was readily segmented into matrix cells and precipitates, allowing for a fast and robust analysis. The deviation of the experimental distribution of Zr, Fe and Cu in precipitates of Z and 1:5 phase from their corresponding random distributions was identified. The method to quantify these deviations is discussed in the next chapter.

The composition space-based approach was not able to capture the accurate number of phases in the Ni-based superalloy because of the large number of features or chemical species. So, a clustering technique employing the random forests algorithm was devised, executing dimensionality reduction in the composition space before proceeding with clustering. Random forest helps to capture the non-linearity in the data and also provides dimensions that are most important, therefore helping in visualizing and interpreting high dimensional composition space. This additional method makes the composition space-based approach equipped for high throughput APT data analysis.

The methods developed and presented in this chapter depend on the contrast in composition of the various chemical domains in the data. If the difference in compositions between the chemical domains is not pronounced, it will be difficult to cluster them in the composition space even after dimensionality reduction. This may also result in an interface between the chemical domains being identified as a separate cluster/phase in the composition space.

Nevertheless, the developed methodology can readily be applied to a wide variety of microstructures in APT for reproducible and robust analysis.

# Chapter 4

# Machine learning based decomposition of geometrically complex chemical domains

*This chapter is a modified version of my publication [48]. The Fe-doped Sm-Co alloy APT dataset was provided by Leopoldo Molina-Luna, Oliver Gutfleisch, Nikita Polin, and Baptiste Gault.*

## 4.1 Introduction

In this chapter, I extend the aim of quantification of APT data to morphologically complex microstructural features. A rigorous method to extract and segment various chemical domains has already been established in the previous chapter. The individual entities comprising a chemical domain/phase can have an intricate morphology, which may have an impact on the properties of the given material. For example, analyzing triple junctions and in-plane composition fluctuations in grain boundaries forming convoluted networks as in [84] gives additional insight from a materials design point of view than extracting just the average composition of the grain boundaries. Geometrically complicated morphology of microstructural regions of interest poses a challenge for human-guided quantification methods, both from temporal efficiency and reproducibility perspectives.

In recent years, machine learning algorithms have been used more often in the analysis of APT data, with the aim of enhancing automation, reliability, and efficiency of the analysis process. Madireddy et al. presented a deep learning-based edge detection to identify the interface between a matrix and a precipitate phase [40]. Although this technique offers a scalable (high-throughput) alternative to iso-concentration surfaces, it does not provide any information about the shape and geometry of the precipitate phase. Peng et al. quantified segregation at grain boundaries through a clustering algorithm applied on spatial coordinates of atoms [85], followed by calculating composition and thickness fluctuations. Zhou et al. introduced a more precise approach to identify junctions and straight segments in connected interface networks, notably grain boundaries, via deep neural-network image recognition in 2D projections [46]. The approach works best for columnar grains projected into their basal plane. An extension to interface networks with random orientation, which lack a common projection plane, is unfortunately not straightforward.

Here, I introduce and showcase the working of an APT data analysis pipeline that detects phases/chemical domains in the data and then disentangles geometrically complex domains into simpler ones. The pipeline uses the clustering in composition approach for phase detection and a U-Net-based model for geometry simplification. Datasets containing microstructural features like entangled planar precipitates/phases, decorated grain boundaries, or decorated inter-phase boundaries can be

readily analyzed using the methodology. The algorithm disentangles the geometry of the features into plate-like structures, enabling meaningful compositional and morphology quantification. Several Fe-doped Sm-Co alloy APT datasets are analyzed in this chapter using the new workflow. This evaluation showcases both the automation of this process and the previously introduced method based on clustering in composition space.

## 4.2 Materials and Methods

### 4.2.1 Experimental data

In Chapter 3, the Fe-doped Sm-Co permanent magnet was introduced. This alloy is subsequently utilized as a case study in the present discussion. Its high-temperature performance and corrosion properties make it an ideal candidate for automotive and aeronautical applications. The spatial distribution of various phases and their composition controls the magnetic properties of this material. In each grain of the alloy, the $Sm_2Co_{17}$ (2:17) or the matrix phase is separated by a network of $Sm(CoCu)_5$ Cu-rich(1:5) phase. The plate-like Zr-rich(Z) phase, which is perpendicular to the c-axis of the 2:17 phase, intersects this microstructure. The interplay between these 3 phases has a crucial part to play in controlling the coercivity of the magnet [86].

Using Kerr microscopy as shown in [86], small regions of lower coercivity compared to the surrounding region were observed near the grain boundaries of this alloy. To find correlations between microstructure and the difference in the magnetic properties, it is important to compare and quantify the microstructure and composition of each phase in 2 regions with different properties. Therefore, three APT samples are each extracted from two regions: one of lower coercivity and the adjacent region of higher coercivity. The APT datasets corresponding to lower coercivity are named L1, L2, and L3 in the following; on the other hand, the ones from higher coercivity are called H1, H2, and H3.

### 4.2.2 APT data analysis pipeline

Figure 4.1 shows the current overarching workflow to classify different chemical domains in any APT data and to reduce morphologically complex precipitates into simpler geometries. The spatial coordinates (x,y,z), mass-to-charge state, and ranges defining the elemental or molecular nature of each ion are extracted from AP Suite, a proprietary software used to reconstruct and analyze APT data (POS and RRNG files) [87]. Further, different chemical domains are extracted from the reconstructed data using clustering in composition space, which is discussed in detail in the previous chapter in Section **Classification of chemical domains: clustering in composition space**. After a DBSCAN-based post-processing step, the morphology of the individual chemical domains is considered. If the APT reconstruction contains precipitates with a complex morphology they are disentangled into plate-like simpler structures using a local PCA approach or a U-Net model, which are described below in Section **Decomposing geometrically complex precipitates**. The plate-like precipitates or structures can be further analyzed by calculating the in-plane composition and thickness fluctuations. Finally, descriptors corresponding to composition, orientation, and thickness for each phase's precipitates are stored systematically using the hdf5 file format [88]. All codes for reading and analyzing APT data are developed in Python, in a Jupyter environment, an open-source web-based tool used for code development and data visualization [89]. ParaView, an open-source scientific visualization tool, is used for additional analysis and visualization [90]. The code for the demonstrated workflow presented in this chapter is accessible at https://github.com/Alaukiksaxena/APT_Machine_learning.

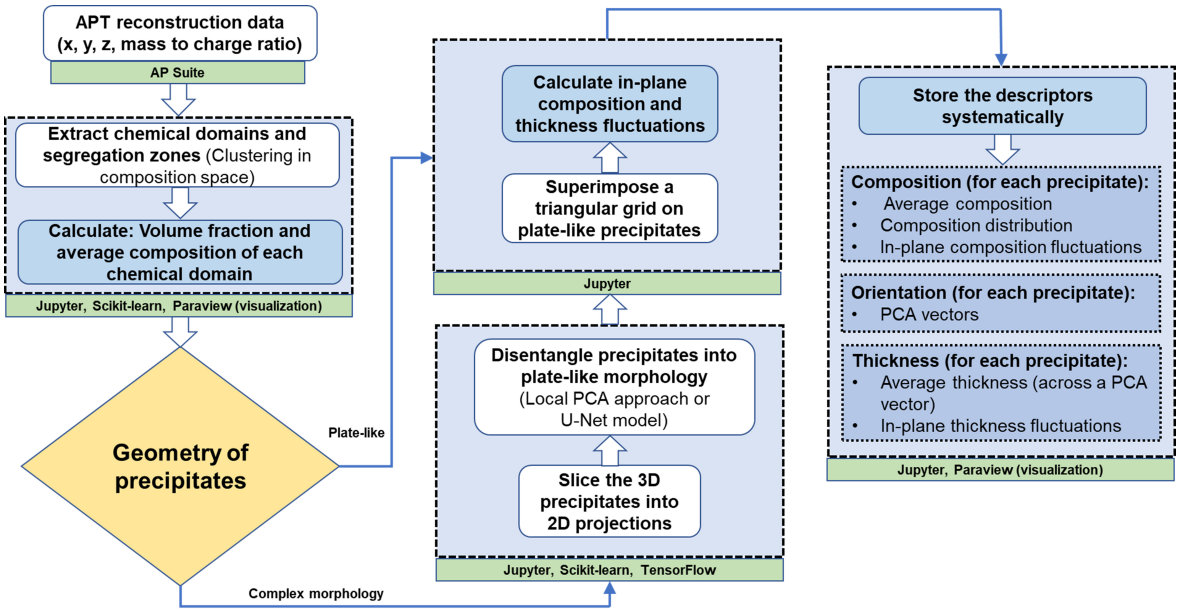



FIGURE 4.1: The steps to quantify a 3D microstructure obtained from an APT reconstruction are summarised in the flow chart. The tools and libraries employed in each step are mentioned in the green boxes.

### Decomposing geometrically complex precipitates

**Local principal component analysis(PCA)-based method**: In APT datasets, especially while studying nano-grain structures, we often encounter quasi-planar interfaces or interphase boundaries that cross each other forming triple or quadruple junctions. This geometric complexity makes the quantification of such microstructures cumbersome and challenging. In the Sm-Co alloy under investigation, the Cu-rich 1:5 phase has precipitates with complex geometry that do not cross the entire field of view, making the approach of Zhou et al. [46] discussed in Section 4.1 impossible to use. One such precipitate of 1:5 phase is encircled in Figure 3.1(E) of the last chapter. To separate complex precipitates into plate-like substructures visible to the human eye, we propose a local PCA-based method [91]. This method is applied to all identified precipitates to find plate-like structures automatically.

This method is demonstrated on a 1:5 phase precipitate, as shown in Figure 4.2(A), with an entangled morphology. First, PCA is performed on the spatial coordinates of voxel centroids belonging to the chosen DBSCAN cluster (or precipitate), and all the voxel centroids are transformed to the PCA coordinate system (PC1, PC2, PC3), Figure 4.2(B). For visualisation a 2D projection of the 3D precipitate in the PC3 direction is shown. Next, the precipitate is divided into $N_{\text{cut}}$ slices, such that each slice is 4nm thick, shown by the dotted lines in the figure, along the PC1 direction. Each slice is projected in the PC1 direction, i.e., to the PC2-PC3 plane, Figure 4.2(B). The slicing process is repeated in the PC2 and PC3 directions to generate 2D projections in the respective perpendicular planes. The created 2D projections are shown in Figure 4.2(C) and will be used to identify the junctions connecting the ribbon-like features in the projection plane.

For this, the local PCA approach is applied to each of the 2D projections or slices. This method helps in the classification of voxel centroids (data points in the 2D projection) to be in junction-like or non-junction-like environments, Figure 4.2(D) to (F). The local PCA algorithm is summarized in Figures 4.3(A) to (D). As shown in Figure 4.3(A), for each data point in a 2D projection (e.g., the orange point), we find all neighborhood points lying within a radius $r$ (green). Subsequently, from the PCA of the distance vectors, the direction of maximum variance is identified, $PC1_{center}$, Figure 4.3(B). The radius $r$ is provided as a hyperparameter for each precipitate. The $PC1_{center}$ direction is then

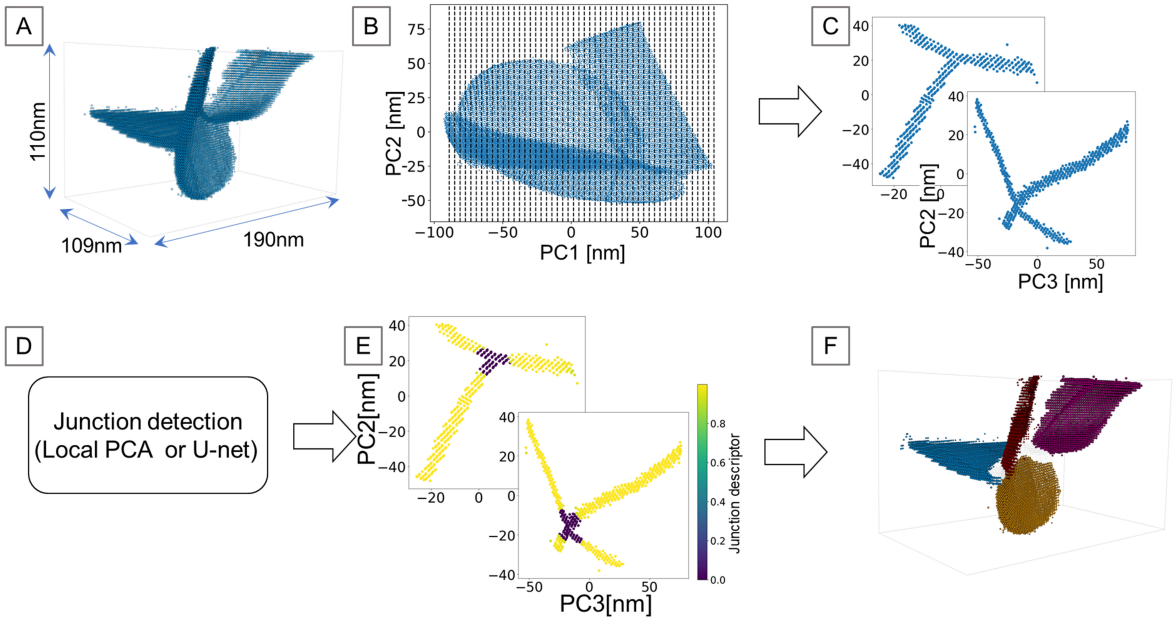



FIGURE 4.2: Workflow for precipitate disentanglement via junction detection and removal. **(A)** Voxel centroids corresponding to a 1:5 phase precipitate (H3 dataset) having four plate-like structures meeting at a junction, also shown in Figure 3.1(E). **(B)** The voxel centroids are transformed to the PCA coordinate system and divided into $N_{\text{cut}}$ slices along PC1 (direction of maximum data variance). For visualisation a 2D projection of the 3D precipitate in PC3 direction is shown. **(C)** 2D projections for slices of voxel centroids in the plane perpendicular to the cutting direction. **(D)** and **(E)** Local PCA approach or U-Net model trained on similar data is used to detect junctions in the 2D projections. **(F)** All the voxel centroids that are labeled as plate-like back-projected to 3D space where they are clustered using DBSCAN.

compared to the corresponding one, $PC1_{neigh}$, for all the data points in the neighborhood (green) by computing the dot product ($PC1_{center} \cdot PC1_{neigh_i}$ where i is a given neighborhood data point). If the selected data point (orange) lies in a quasi-plate-like structure then most of the $PC1_{neigh}$ directions will point in the same direction as $PC1_{center}$ and the average of their dot products will be close to one (threshold for not-junction: average dot product $> 0.8$). However, if the orange data point lies near the junction then most of the $PC1_{neigh}$ directions will not point along $PC1_{center}$ and the average of their dot product is lower (threshold for junction: average dot product $\leq 0.8$), Figure 4.3(C). In this way, the local PCA approach, which is applied to all the data points in a 2-D slice, is able to classify the data points that lie at the junction of the precipitates, Figure 4.3(D). All the junction data points from each 2-D slice are stored in a list and tracked back to the original points (or voxel centroids) in the 3-D spatial coordinate system. Finally, the junction centroids in 3D space are removed from the set of all the voxel centroids pertaining to a particular precipitate, separating the quasi-plate-like structures. These structures can now be identified as separate clusters by applying DBSCAN since the connecting junction between them has been removed, Figure 4.2(F).

**U-Net based approach**:

In the local PCA-based approach, the accuracy of the classification of data points as junction depends on the hyperparameter $r$, which has to be fed for each precipitate, shown in Figure 4.3(A). For example, if the radius $r$ is larger than the precipitate the local PCA approach will not be able to find the data points at the junction. To circumvent this problem, we have developed a more robust approach based on supervised machine learning (U-Net) and image augmentation. A U-Net is a convolutional neural network (CNN) based architecture, which is used in computer vision problems, especially biomedical image data, for semantic segmentation [59]. Semantic segmentation means that the deep learning algorithm predicts the class for each pixel in a given image.

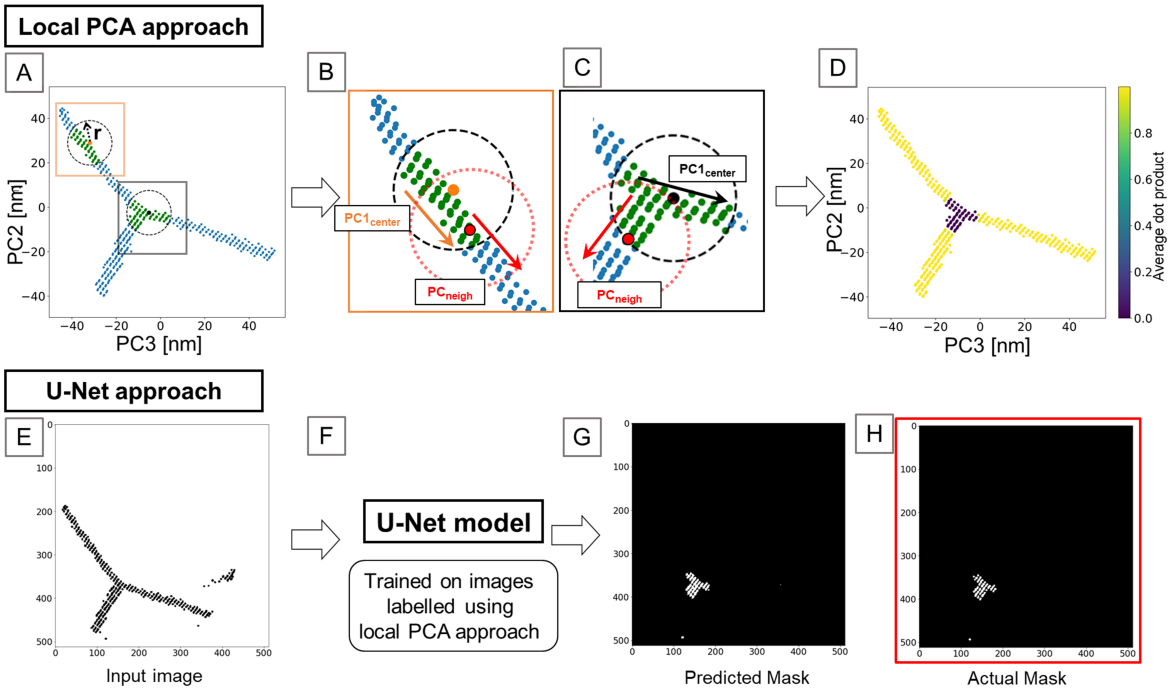



FIGURE 4.3: **Junction detection. (A)** A 2D projection of the voxel centroids in a slice extracted from a 3D precipitate (H3 dataset). The orange and the black box highlight a plate-like and junction region, respectively. The neighborhood (green) within a radius $r$ around a chosen point (orange) is selected within the orange box. **(B)** Magnified region away from the junction. The direction of maximum variance, $PC1_{center}$ and $PC1_{neigh}$ corresponding to the neighborhood of the orange and red points within radius $r$, respectively. The red point is chosen from the neighborhood of the orange point. **(C)** Magnified region near the junction. **(D)** Average dot product mapped on the 2D projection delineating the junction region. **(E) to (H)** A U-Net model trained on local PCA approach labeled images is applied on a previously unseen input image to predict the junction region or mask. The actual or true mask is also shown.

First, the local PCA approach is applied to 1:5 phase precipitates extracted from a given APT dataset. This process created 500 2D projections or slices. The data points in the projections are already labeled as junction or not junction using the local PCA approach. All the projections are converted to gray-scale images of shape $512 \times 512 \times 1$ and based on the label information a mask depicting the junction in each image is also extracted with shape $512 \times 512 \times 2$, where the last 2 channels correspond to the classes: junction and not junction.

To increase the diversity of the training data, image augmentation steps like rotation, translation, flipping and resizing are applied to the original images to create a pool of 3000 training images and their corresponding masks. These training images ($512 \times 512 \times 1$) and the masks ($512 \times 512 \times 2$) are used to train a U-Net. The trained model, as shown in Figures 4.3(E) to (G), takes an image corresponding to a projection as input ($512 \times 512 \times 1$) and predicts a mask of shape $512 \times 512 \times 2$, where the last channel contains the probability of a pixel to be a junction. Based on the predicted mask, the data points in the original projection/slices are classified as junction or not.

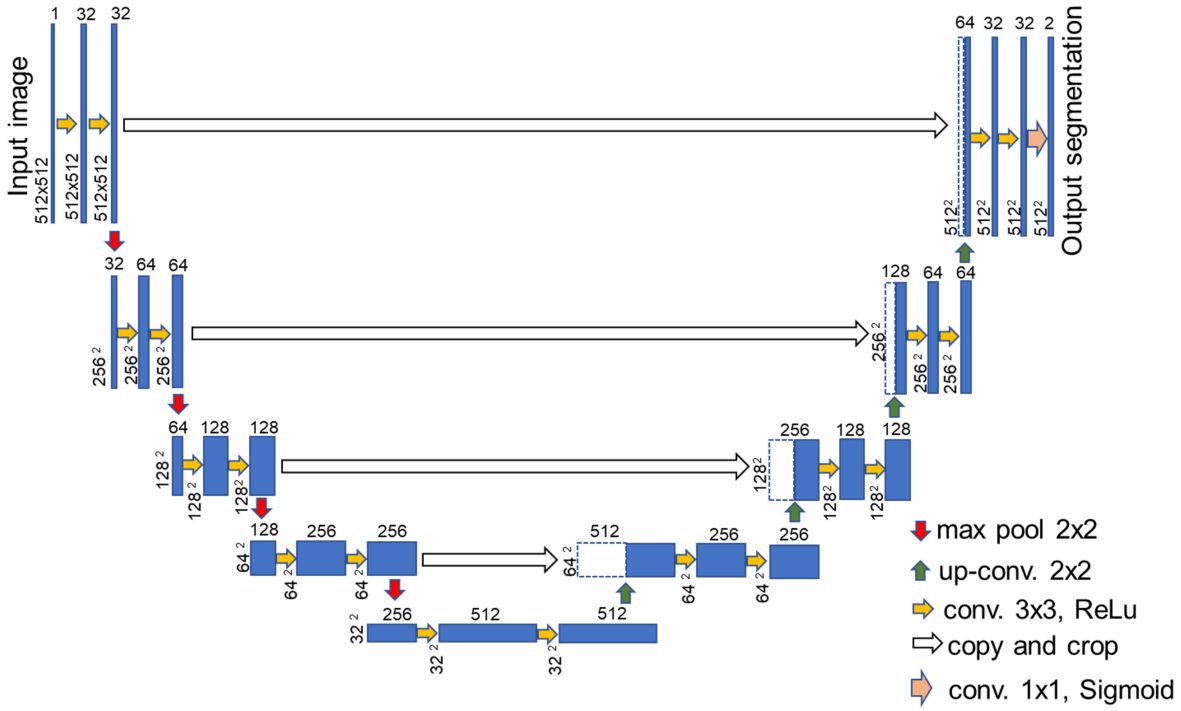



FIGURE 4.4: The architecture of U-Net image segmentation model used in this study. Categorical cross-entropy is used as the cost function for the model which is optimized using Adam optimizer (learning rate = $e^{-5}$)

The U-Net based model is more robust and is directly applied to the other samples of the same material without the need to tune the parameter $r$ for each precipitate. The local PCA-based approach is semi-automatic and can be tuned according to the precipitate size. However, the U-Net-based approach is fully automatic irrespective of the precipitate size. Nevertheless, retraining of the U-Net model might be required when applying the same model to other material datasets, in which the geometry of precipitates is different. The details of the U-Net architecture used in this chapter are given in Figure 4.4.

## 4.3 Results and discussion

The clustering in the composition space approach, which is the first step in the developed pipeline, is applied to all six APT datasets of Sm-Co alloy. The voxels in each dataset are segmented into

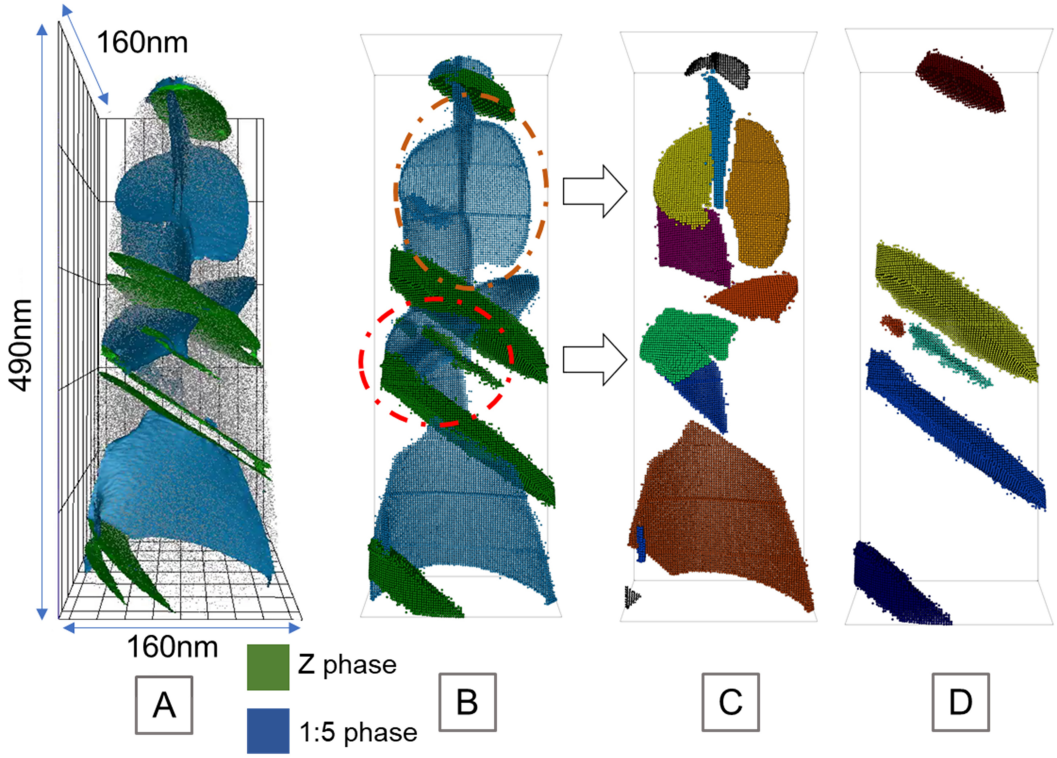



FIGURE 4.5: **(A)** 6 at% Zr (green: Z phase) and 12 at% Cu (blue: 1:5 phase) isosurfaces in H3 dataset extracted using AP suite. **(B)** The voxel centroids lying in the Z phase (green) and 1:5 phase (blue) precipitates extracted using clustering in composition space. The 1:5 phase precipitates with complex morphology are highlighted. **(C)** Segmented 1:5 precipitates. Complex morphology precipitates are separated into plate-like structures. **(D)** Segmented precipitates of the Z phase.

TABLE 4.1: Average volume fraction, thickness, and composition of each phase of six Fe-doped Sm-Co alloy APT data sets.

| | | | Composition (at.%) | | | | |
|---|---|---|---|---|---|---|---|
| Phase | Volume Fraction (%) | Thickness (nm) | Sm | Co | Fe | Cu | Zr |
| High coercivity (H1, H2, H3) | | | | | | | |
| 2:17 | 78.0±1.6 | - | 11.0±0.1 | 59.3±0.2 | 27.2±0.1 | 2.1±0.2 | 0.4±0.1 |
| 1:5 | 6.7±0.5 | 11.7±1.5 | 14.3±0.3 | 48.4±1.0 | 16.5±1.1 | 20.3±1.8 | 0.4±0.1 |
| Z | 15.2±1.1 | 11.7±4.6 | 10.2±0.8 | 62.0±1.9 | 15.4±1.5 | 3.9±0.7 | 8.5±1.3 |
| total | - | - | 11.3±0.1 | 58.1±0.2 | 24.7±0.2 | 4.5±0.1 | 1.3±0.1 |
| Low coercivity (L1, L2, L3) | | | | | | | |
| 2:17 | 65.1±2.8 | - | 11.0±0.1 | 59.2±0.1 | 27.7±0.1 | 1.6±0.1 | 0.4±0.1 |
| 1:5 | 14.8±0.9 | 11.0±2.7 | 15.0±0.4 | 52.3±0.6 | 17.0±0.9 | 15.3±1.1 | 0.4±0.1 |
| Z | 20.1±1.9 | 10.8±2.8 | 11.7±0.5 | 60.7±0.7 | 16.4±1.1 | 3.8±0.4 | 7.4±1.2 |
| total | - | - | 12.3±0.2 | 57.4±0.2 | 22.9±0.5 | 5.8±0.4 | 1.5±0.2 |

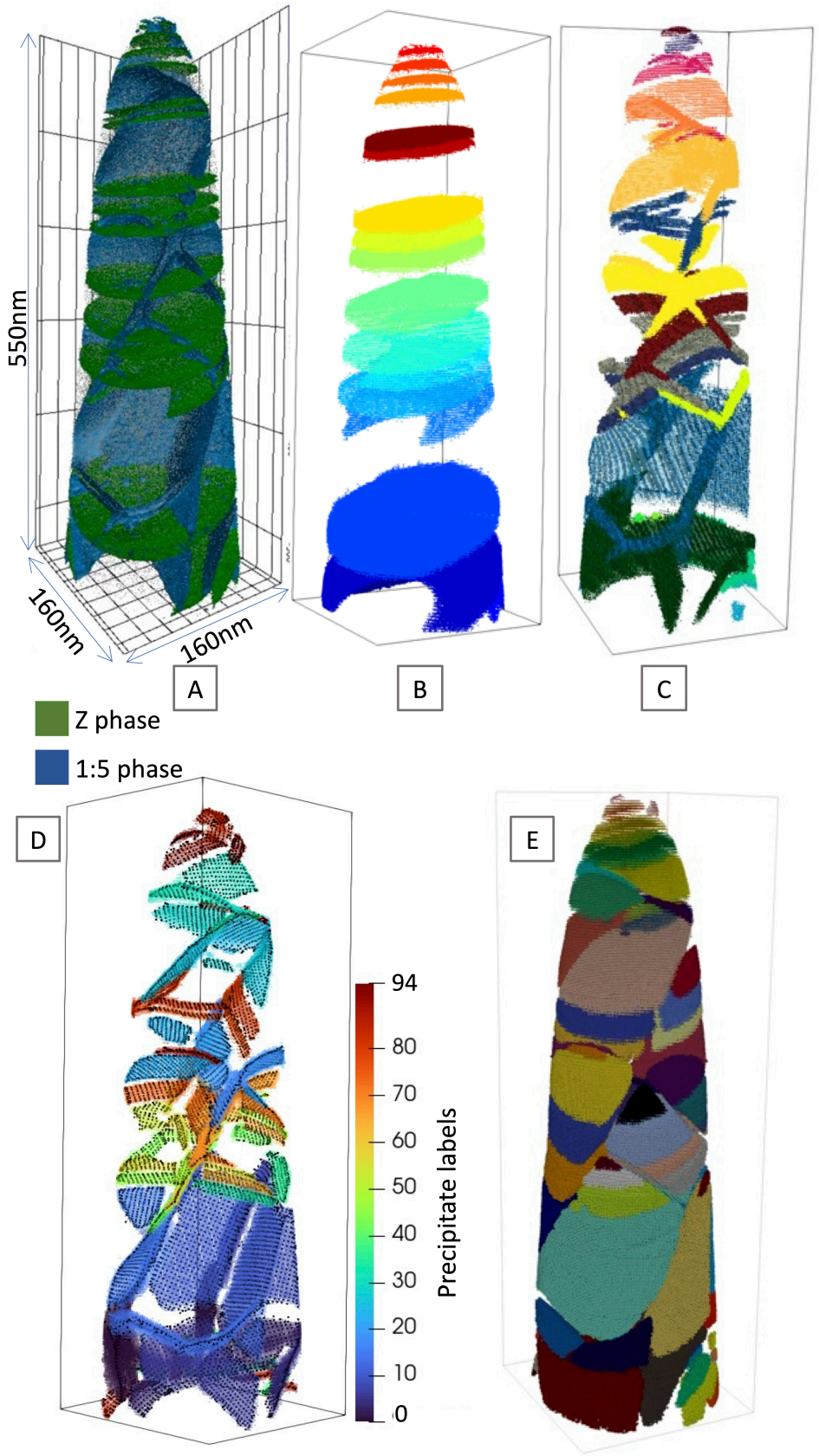



FIGURE 4.6: **(A)** 6 at% Zr (green: Z phase) and 12 at% Cu (blue: 1:5 phase) isosurfaces in L3 dataset extracted using AP suite. **(B) and (C)** Segmented precipitates of the Z phase and 1:5 phase after applying clustering in the composition space approach. **(D)** Precipitates of 1:5 phase with complex morphology are reduced to plate-like structures. **(E)** Segmented matrix cells after removing 1:5 and Z phase.

three phases, namely: 2:17, 1:5, and Z phase. The composition space corresponding to the H3 dataset is shown in the previous chapter in Figure 3.1 (C). The segmented microstructure for H3 is shown in Figure 4.5(B), which is visually identical to the isosurface results in Figure 4.5(A). Similarly, the segmented Z phase, 1:5 and 2:17 phase for L3 dataset are shown in Figure 4.6 (B), (C), and (E) respectively. They are also visually identical to the corresponding isosurfaces shown in Figure 4.6 (A). L3 has a much more complicated microstructure than the H3 dataset. However, the composition space approach works out of the box without changing any hyperparameter to segment the complicated microstructure into respective phases and precipitates, and readily identifies 33 precipitates of the 1:5 phase and 18 precipitates of the Z phase within a few minutes for the L3 dataset.

After getting the voxels of the respective phases in each sample, phase volume fraction and composition for each chemical species averaged over the six datasets are computed as shown Table 4.1 (details of calculations mentioned in section 3.3.2). On average, the datasets from the low coercivity region have approximately 5% more Z phase and 8% more 1:5 phase compared to the ones with high coercivity. Considering all the phases (mentioned as "total" in Table 4.1), the datasets with low coercivity have higher Sm composition, which is also the reason why the regions with low coercivity appear brighter in the backscattered electron scanning micrographs [86], compared to the high coercivity surrounding region. Also, the 1:5 phase of the high coercivity region has $5\pm2.9$ at.% more Cu and $3.9\pm1.6$ at.% less Co compared to 1:5 phase of the low coercivity.

Subsequently, the trained U-Net model is applied on each segmented precipitate of each phase to remove the junction voxels. Thereafter, the quasi-plate-like structures are retrieved if the precipitate has a complex morphology. Figure 4.5(C) shows 1:5 precipitates in the H3 dataset, which are separated into plate-like structures for further analysis. The Z-phase precipitates are also shown in panel (D) of the figure. Since they already have a planar morphology without any junctions, they are not modified by the U-Net-based approach. The more complicated network of 1:5 phase in the L3 dataset is already segmented into 33 separate entities/precipitates by composition space-based approach (see Figure 4.6(C)). These precipitates are further simplified into 95 plate-like features as shown 4.6(D). The U-Net model was trained using a dataset derived from 2D images or projections of a 1:5 precipitate from the H3 dataset, which were enhanced through data augmentation techniques. For training, masks or labels indicating junction regions in these images were generated using the local PCA approach. Despite the training on this specific dataset, the U-Net model was capable of automatically and accurately identifying junctions in the more complex 1:5 precipitates of the L3 dataset. Furthermore, the model aided in reducing these complex structures into planar formations.

### 4.3.1 In-plane composition and thickness fluctuations

After segmenting the Z-phase precipitates and reducing each 1:5 phase precipitates into planar structures, in-plane composition and thickness fluctuations for each feature are calculated for all the given datasets. This analysis helps understand, for example, the composition fluctuations in precipitates of different phases, as shown in Figure 3.6 for H3 dataset.

The algorithm of Peng et al. is modified to find the in-plane composition and thickness fluctuations in the plate-like precipitates [85]. The first step is to apply a PCA transformation on the voxel centroids of a given quasi-planar precipitate. Next, a regular 2-D triangular grid of size 4 nm is superimposed on the precipitate in the plane formed by the two maximum variance PCA directions. The grid triangles not containing voxel centroids are removed. This is followed by moving each node of the trimmed grid to the center of mass of the voxel centroids located in the cuboidal region overlayed at that node, shown in Figure 4.7 (A). We use voxel centroids of the precipitates for analysis, while the original algorithm directly uses atoms at the interface to be analyzed. Since some of the precipitates in this

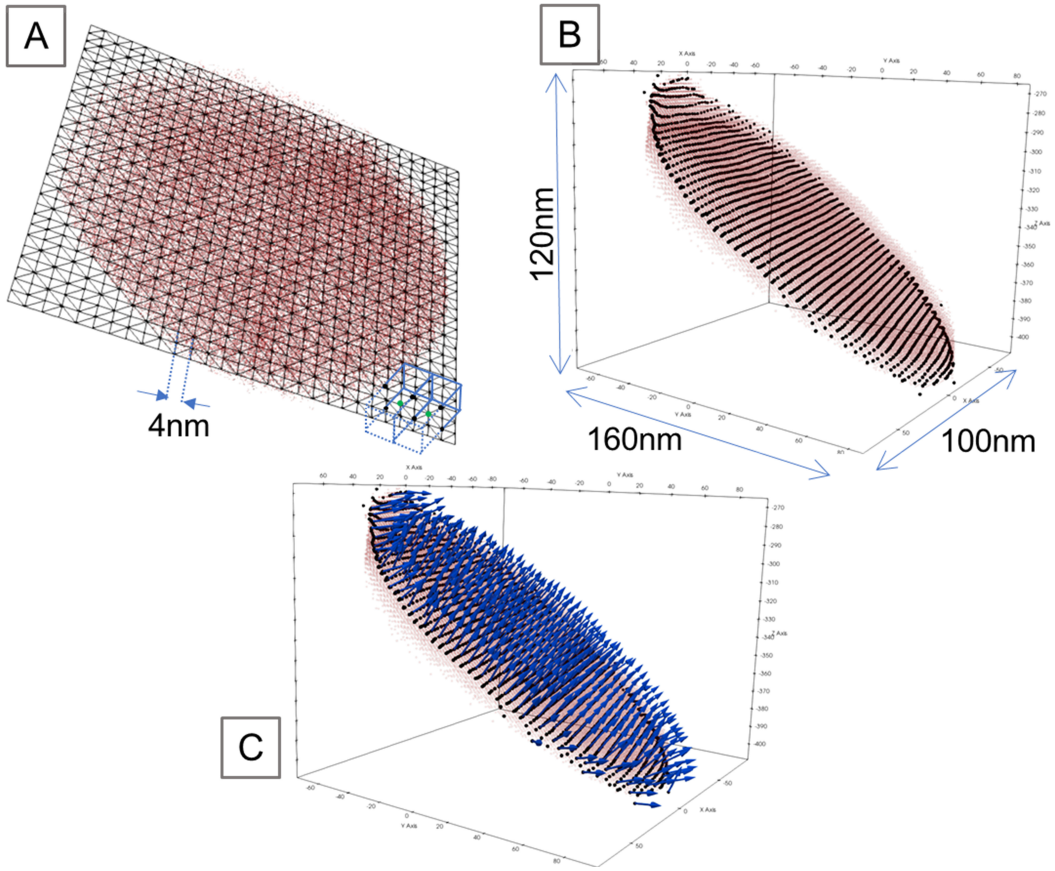



FIGURE 4.7: **(A)** Voxel-centroids that comprise the plate-like precipitate are transformed into PCA coordinate system. A 2D triangular grid is superimposed on the precipitate in the plane formed by the PCA vectors with maximum standard deviation. The grid triangles not containing voxel centroids are removed. **(B)** Each node of the grid is moved to the center of mass of the voxel-centroids contained in the cuboid region centered at a given node. The cuboidal region for two green nodes is shown in panel **(A)**. As a result the grid takes the shape of the precipitate. **(C)** At each node the normal to the grid surface is evaluated and atoms present in a cylindrical region of interest (ROI) along the normal direction are selected. Further, average composition, thickness and 1D composition profiles are calculated along each ROI which enables the investigation of in-plane composition and thickness fluctuations.

case study have a million atoms or more, the choice of voxel centroids makes it faster. As a result of moving the nodes, the trimmed grid takes the shape of the precipitate, Figure 4.7 (B) and (C). The geometry of the precipitate is still preserved if voxel centroids are used for the analysis.

After obtaining the grid that conforms with the shape of the precipitate, the local surface orientation (normal vector) at each node is defined from the average of the adjacent triangles (see Figure 4.7 (C)). This is followed by selecting atoms in a cylindrical region of interest (ROI) at each node with the axis of the cylinder along the node's normal. To speed up the atom search, we exploit the space partitioning provided by the voxels. First, relevant precipitate voxels are identified, whose centroids fall inside a cylindrical ROI with a radius, $r_{out}(= 4\,\text{nm})$, and length $L_{out}(= 34\,\text{nm})$. From the selected voxels, the atoms within a cylindrical ROI of dimensions, $r_{in}(= 2\,\text{nm})$, and length $L_{in}(= 30\,\text{nm})$, are chosen for the calculation of in-plane thickness and composition fluctuations.

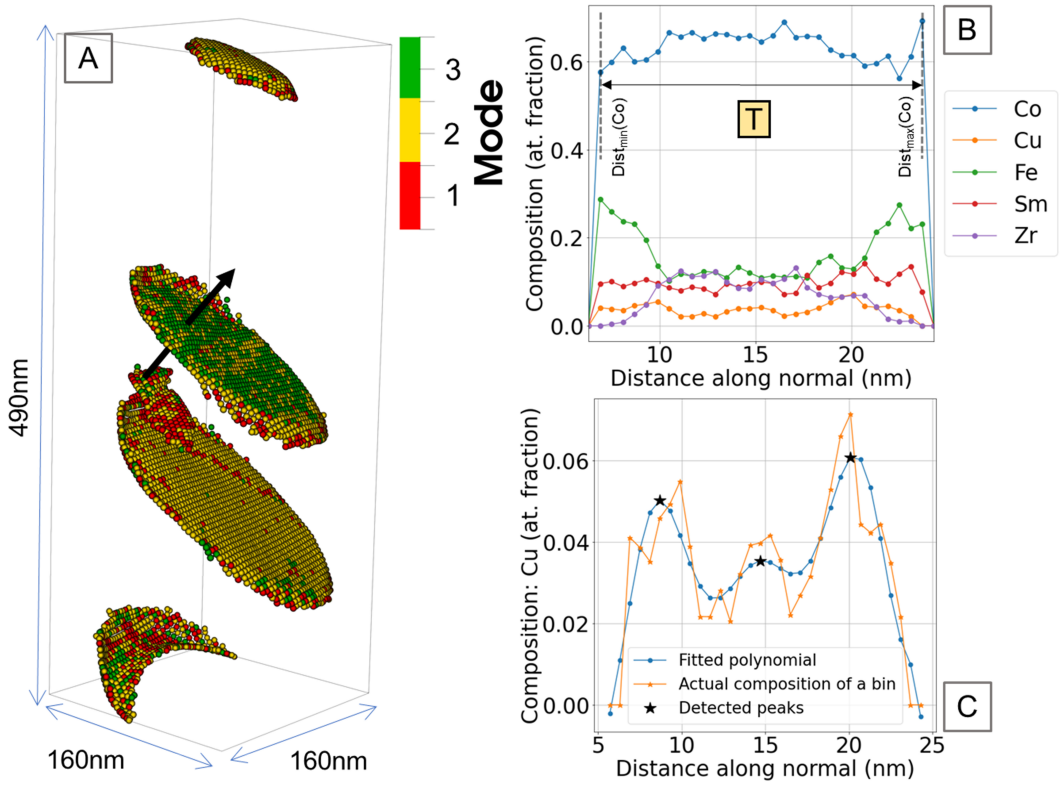



FIGURE 4.8: **(A)** The nodes in a triangular grid superimposed on the Z phase precipitates (H3 dataset). The color of each node pertains to the number of modes or peaks in the 1D composition profile of Cu at that particular node. **(B)** 1D composition profile of Co, Cu, Fe, Sm, and Zr along a cylindrical ROI around the black arrow (normal at a node), crossing through the precipitate shown in panel **(A)**. Here, only the atoms within the precipitate are considered to plot the 1D composition profiles. Also, the composition profiles while considering all atoms in the vicinity of the precipitate as well are shown in the Figure 4.9. The thickness (T) of the precipitate along the ROI is estimated by calculating the maximum distance between the bins that have non-zero Co composition.**(C)** 1D composition (at. fraction) profile (orange) of Cu along the same ROI. A six-degree polynomial (blue) is fitted on the 1D composition profile. Local maxima or peaks (black) are estimated in the fitted polynomial to approximate the local segregation zones of Cu across the precipitate.

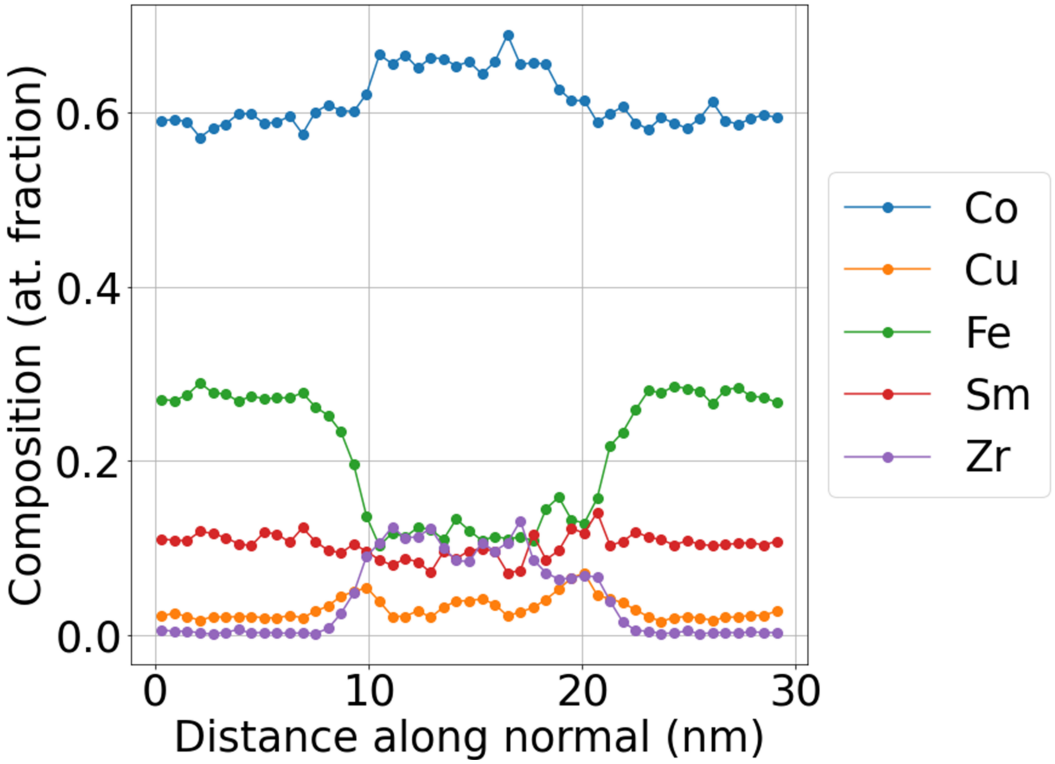



FIGURE 4.9: 1D composition profile of Co, Cu, Fe, Sm, and Zr along a cylindrical ROI around the black arrow shown in Figure 4.8(A).Here, all the atoms in the vicinity of the precipitate are considered as well.

Each cylindrical ROI is divided into bins along the cylinder axis and the composition of each bin is calculated. I arbitrarily selected a bin width of $L_{bin}(= 0.6nm)$ as it provided a good compromise of limited statistical fluctuation and details of the feature of interest. The 1D composition profile for all the elements, obtained at a certain node in a precipitate of the Z phase (H3 dataset) along the normal shown in the Figure 4.8 (A), is plotted in Figure 4.8 (B). This plot considers only the atoms within the precipitate and not the surrounding atoms. Choosing the atoms in this way makes the calculations faster, especially when many such 1D composition profiles are calculated for hundreds of automatically placed ROIs. Complete 1D compostion profiles while considering even the atoms outside the given precipitate for the normal shown in Figure 4.8 (A) is provided in Figure 4.9.

To analyze the clustering or segregation of the chemical species across the quasi-planar precipitates, the 1D composition profiles are studied in more detail. To exemplify our approach, Figure 4.8(C) presents a 1D composition profile of Cu across the aforementioned node in the Z-phase precipitate. A six-degree polynomial is fitted on the 1D composition profile. Subsequently, the number and location of local maxima in the fitted polynomial are obtained using find_peaks, a tool implemented in open source scipy library [92]. Figure 4.8(A) shows the fluctuations in the number of modes (or peaks) in 1D Cu composition profiles of all the Z phase precipitates in H3 dataset. The precipitates exhibit 1, 2, or 3 modes in their composition profiles, suggesting that Cu is either uniformly distributed, segregated at the edges, or additionally segregated at the middle of the precipitates, respectively.

Based on the 1D composition profiles at each node, it is possible to reveal possible in-plane composition fluctuations, as shown in Figure 4.10 for Zr (A and B) and Cu (C and D) for precipitates of Z (A and C) and 1:5 phase(B and D) of the H3 dataset. Since the composition of Co is significantly higher than any other element in both Z and 1:5 phases (refer Table 4.1) and also its amount is higher in the Z phase and lower in the 1:5 phase compared to the matrix (2:17 phase), it is the best choice to mark the start and end of a precipitate. The in-plane fluctuations are calculated by taking the average

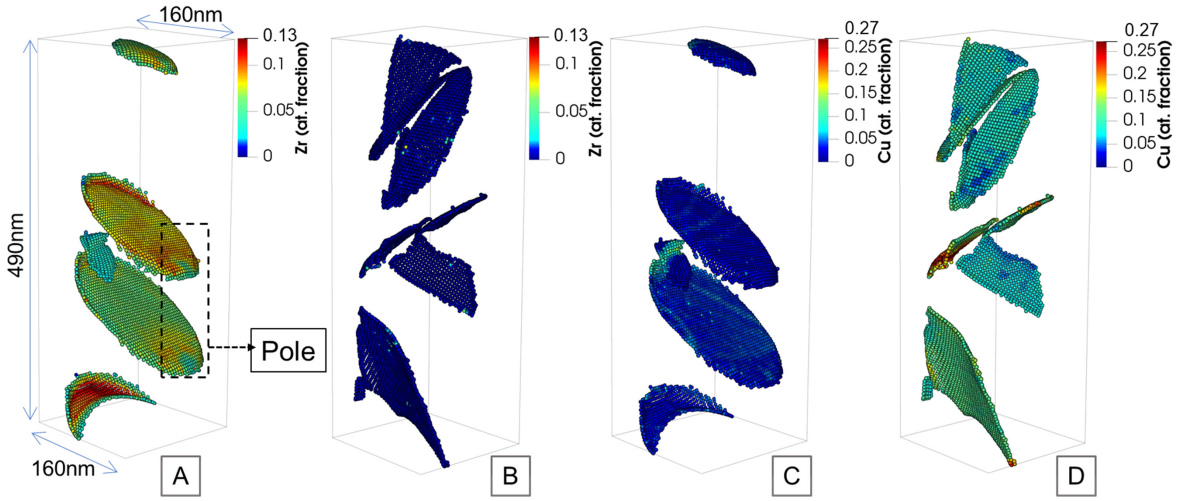



FIGURE 4.10: **In-plane composition fluctuations**: The nodes associated with 2D grids fitted on plate-like precipitates in H3 dataset. The color coding corresponds to the average composition (at. fraction) of Zr (**A**, **B**) and Cu (**C**, **D**) at each node in Z phase(**A**, **C**) and 1:5 phase (**B**, **D**). Further, in panel **(A)**, Zr is depleted in a circular region in two precipitates, marked with a rectangular box. This can be a pole in the dataset, which is an APT artifact.

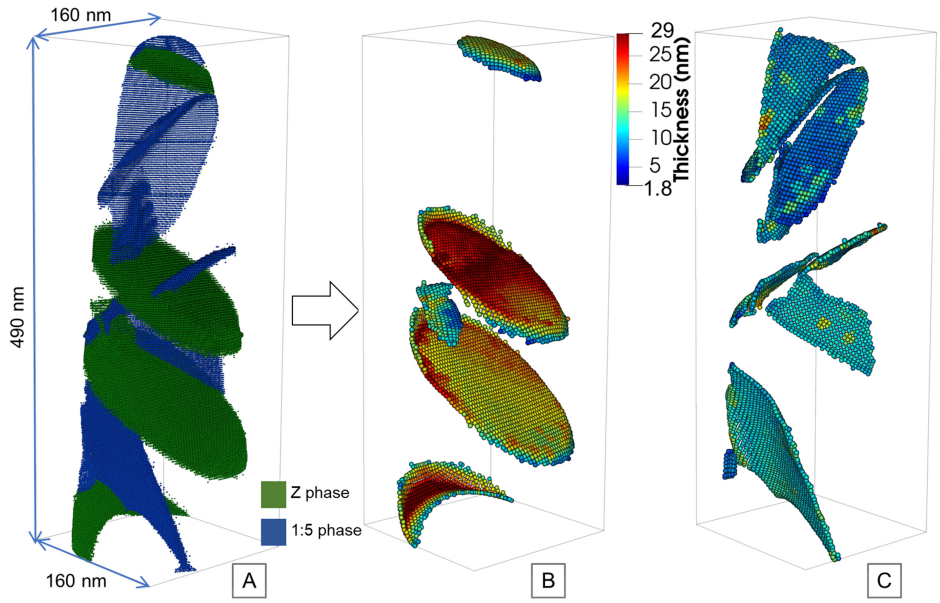



FIGURE 4.11: **(A)** Voxel-centroids of the Z (green) and 1:5 (blue) phase in H3 dataset are shown. Nodes of the triangular grid fitted on each plate like precipitate with thickness (nm) of the precipitate at each node is illustrated in **(B)** and **(C)** for Z and 1:5 phase respectively

of an element's composition over the length of the ROI where Co composition is not zero. As shown in Figure 4.8 (B), the average is calculated over the bins lying between $Dist_{max}(Co)$ and $Dist_{min}(Co)$. This line of reasoning is also used to calculate the thickness of each precipitate at every node of the grid. Thickness, T = $Dist_{max}(Co) - Dist_{min}(Co)$, of the Z precipitate at the node, is shown in Figure 4.8(B). Thickness fluctuations for precipitates of Z and 1:5 phase in H3 are summarised in Figure 4.11(B) and (C). The calculation of the average thickness for the precipitates in each phase is done by averaging the thickness measurements taken at every node. The computation of in-plane composition and thickness fluctuations in the precipitates can take up to 30 minutes on a single CPU core, for a fine grid size (4nm), for the H3 dataset.

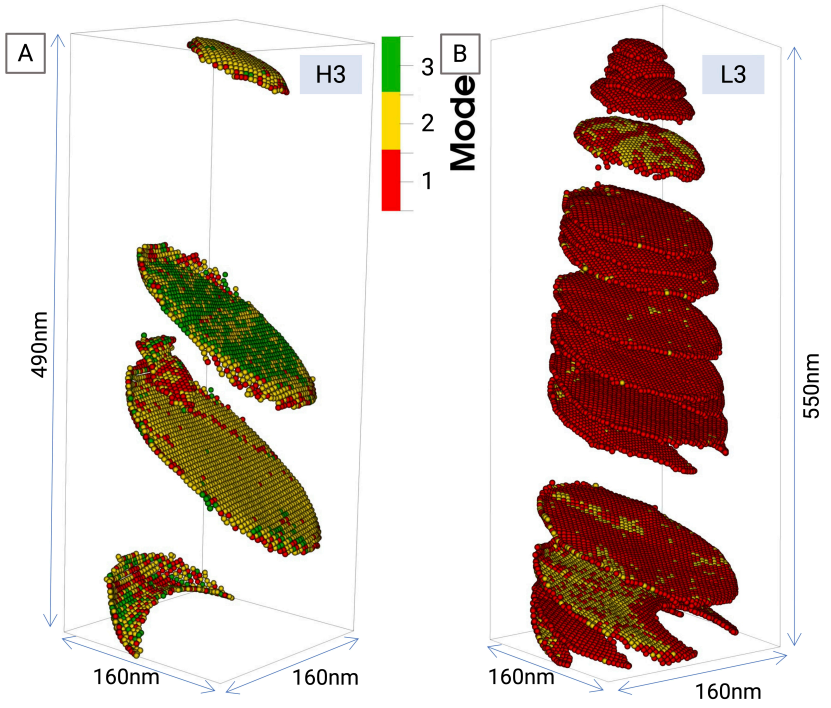



FIGURE 4.12: The nodes of triangular grid fitted on each Z phase precipitate of H3 **(A)** and L3 **(B)** datasets. The number of modes or peaks in the 1D composition profile of Cu is mapped on each node.

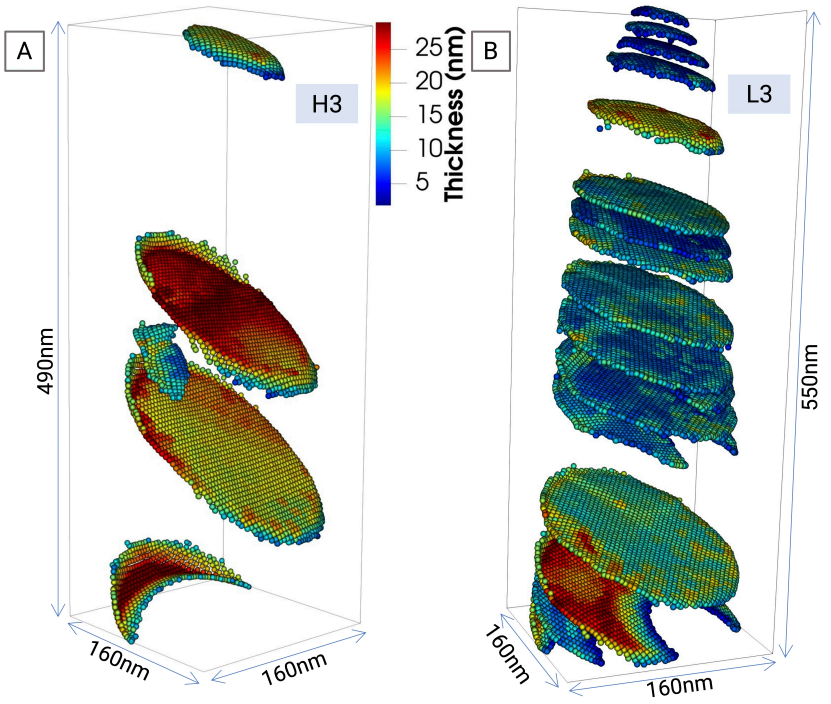



FIGURE 4.13: The nodes of triangular grid fitted on each Z phase precipitate of H3 **(A)** and L3 **(B)** datasets. The thickness of precipitates at each node is mapped on the grid.

The most pronounced difference between the two types of datasets is observed in the fluctuations in the number of peaks/modes in the 1D composition profile of Cu in Z-phase precipitates. Figure 4.12 shows grids fitted on the Z phase of H3 and L3 datasets. The different colors of the grid nodes in the figure depict the number of peaks in the Cu 1D composition profile. Clearly, most of H3 dataset

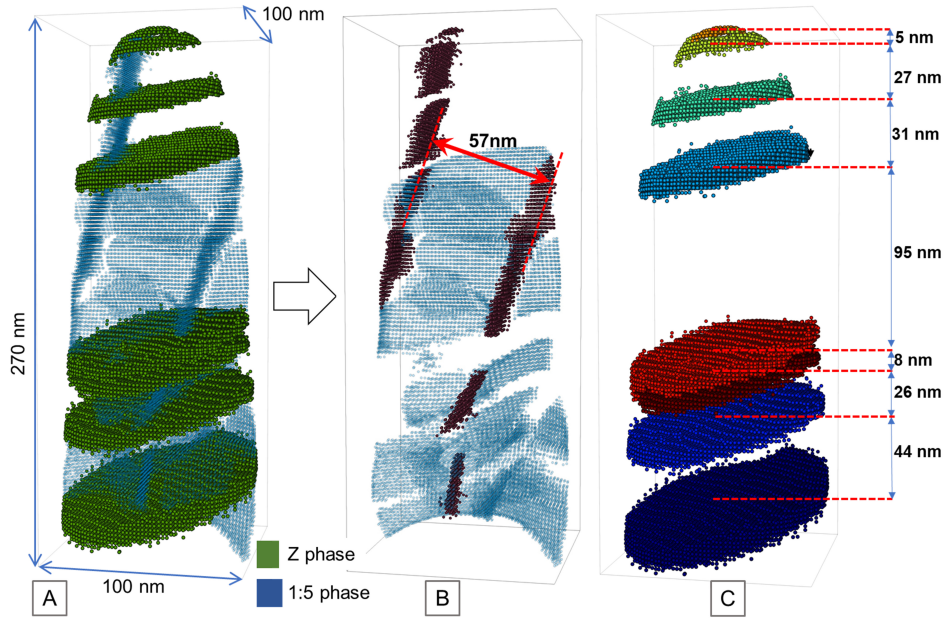



FIGURE 4.14: **(A)** Voxel-centroids of Z and 1:5 phase, green and blue respectively, present in L2 dataset with complex microstructure. **(B)** and **(C)** Sets of parallel 1:5 and Z phase precipitates, respectively, with distances between them.

precipitates have 2 or 3 peaks. On the contrary, for L3 dataset, most of the precipitates have one peak, and a few have two peaks. This difference is observed in all six APT datasets analyzed. On top of this, Figure 4.13 exhibits the thickness fluctuations in the Z phase precipitates of H3 and L3. On correlating this image with Figure 4.12 it can be observed, for both datasets, that the regions in the precipitates that are thicker have a higher possibility of having multiple peaks in the composition profile of Cu.

### 4.3.2 Quantification of the microstructure

To quantify the Sm-Co alloy microstructure, in addition to the thickness, orientation, and spacing between the extracted precipitates have to be calculated. The parameters like the distance between parallel precipitates and the angle between the intersecting ones are used to define a 2D model microstructure of a Sm-Co alloy [93]. Similarly, we have extended this concept to quantify 3D microstructures. To define the orientation of each plate-like precipitate, PCA is carried out for the spatial coordinates of voxel-centroids present in the given precipitate. The direction of minimum variance in the spatial coordinates is then taken as an approximate orientation normal to the precipitate plane. Figure 4.14(A) shows voxel centroids corresponding to Z and 1:5 phase in L2 dataset. The 1:5 phase precipitates with complex morphology in the given microstructure have been reduced to plate-like structures. The angle between each pair of planar precipitates is calculated and a set of parallel precipitates (the angle between the normal vectors is in range $0^\circ$ to $5^\circ$) forming the magnetic domain walls is shown in Figure 4.14(B). Similarly, a set of parallel Z phase precipitates in the same APT sample is shown in Figure 4.14(C). The distance between a pair of parallel precipitates is the distance of the centroid of one of the precipitates from the plane of the other precipitate. The distances between the consecutive Z phase plates, and a set of parallel 1:5 phase precipitates are summarised in Figure 4.14(B) and (C). This computational analysis can be executed in just a few seconds using a single CPU core for the given APT dataset.

Such quantification of the microstructure from APT datasets is directly dependent on the way APT data is reconstructed. In the present study, the APT data has been reconstructed in AP Suite Software 6.1 using voltage reconstruction. The image compression factor (ICF) was set to 1.65 and the radius was adjusted such that the Z-phase precipitates have the form of straight platelets and that these platelets are aligned parallel to each other in the reconstruction as observed in transmission electron microscope (TEM) measurements. The distances between the platelets observed in APT are compatible with those in TEM measurements on the same sample [86]. In the future, the distances extracted from my algorithm could be used to fine-tune APT reconstruction parameters such that they reproduce TEM data quantitatively.

The quantitative analysis of composition and geometric aspects of the chemical domains identified by the 3D segmentation algorithm illustrates opportunities that come with automatic identifications of microstructural features. It is very easy to apply new APT data analysis methods as shown in this section, or adapt already existing ones like the construction of 1D composition profiles or composition fluctuation maps. Thanks to the controlled workflow, this can be done not only for selected regions of interest but automatically for all features identified. This allows for a more comprehensive description of the APT data than what was possible so far.

## 4.4 Conclusion

In this chapter, an APT data analysis pipeline was developed. The pipeline was used to compositionally and geometrically quantify APT datasets corresponding to a Fe-doped Sm-Co magnet in an automatic way. The microstructure of the datasets was segmented using clustering in the composition space approach. Thereby, features like volume fraction and average composition of each phase were extracted for all the given datasets.

Geometrically quantifying the microstructure requires reducing morphologically entangled phases to individual structures. This was achieved here for a particular type, namely agglomerates exhibiting plate-like substructures. The novel local PCA and/or U-Net approaches helped to distinguish junction voxels from those lying in the planar regions. With this labeling, DBSCAN clustering is able to get the separated planar structures. This procedure can efficiently and rigorously segment out predominantly flat subdomains with arbitrary orientations in 3D. Indeed, planar or plate-like morphologies are common microstructural features, for example, segregation at the interface (grain boundary) and interphase boundaries or phases formed in some Al [94], [95] and Ni alloys [96]. Then descriptors such as PCA vectors, their explained variance, average composition, centroid, or the angle of the inclination with the axis of the APT tip, are calculated for each planar substructure. Using the geometrical information, the compositional analysis can be further refined to yield in-plane composition and thickness fluctuations. A method to map the segregation of solute atoms across interfaces has also been developed.

The Sm-Co alloy analyzed here contains regions of low coercivity embedded in grains of higher coercivity. This alloy contains 3 phases that, after heat treatment, self-assemble in such a way that the matrix (2:17) phase is divided into cells by a network of Cu-rich 1:5 phase, and both of them are intersected by Zr-rich plate-like Z phase [86]. Given the complex microstructure of the material, it is difficult to pinpoint the explicit reason for the property differences. Therefore, the APT datasets from these two regions were rigorously analyzed by the developed pipeline. Using the workflow, the three phases, 2:17 phase, Z phase, and 1:5 phase, were correctly identified in all the APT datasets. 1:5 phase precipitates with complex morphologies were reduced to plate-like structures using the U-Net-based approach. The inplane composition and thickness fluctuations for each phase and dataset

were calculated. Two prominent differences were noticed between the datasets from the two regions. Firstly, the 1:5 phase in the high coercivity region is richer in Cu and depleted in Co compared to the lower coercivity region. Secondly, the across-plane composition fluctuation of Cu was very distinct, with enrichment at the edges or also at the center of the Z-phase precipitates in the high coercivity region. Only a few thick Z-phase precipitates exhibited enrichment at the edges in the low coercivity region. When incorporating these distinguishing factors in micromagnetic simulations, they explain the property differences between the two regions [86].

The developed data analysis pipeline can readily be applied to APT datasets containing grain boundaries, phase boundaries, or planar precipitates. Such an analysis will enable a consistent quantification of compositional and structural features in APT datasets.

# Chapter 5

# Analyzing linear features in atom probe tomography datasets using skeletonization

*This chapter is a modified version of my publication [49]. The Ni-based single-crystal superalloy APT dataset was provided by Paraskevas Kontis and Baptiste Gault. The Fe-Mn alloy APT dataset was provided by Alisson Kwiatkowski da Silva and Baptiste Gault.*

## 5.1 Introduction

Dislocations are 1D crystallographic defects that are found in metallic systems [97]. The complex interplay between dislocations and other microstructural elements — such as additional dislocations, grain boundaries, and secondary phases — exerts a profound influence on the macroscopic mechanical properties of the material in question [98] [99]. Dislocations, owing to their associated strain fields, manifest elevated energy states in comparison to the contiguous lattice matrix [98]. This heightened energy environment serves as a preferential site for solute segregation, thereby affecting the local chemical composition and influencing the overall material properties [97][100]. Segregation at dislocations can also lead to the formation of defect phases called linear complexions if the local composition and the stress field associated with the dislocation stabilize them [101]. Overall, the segregation to these line defects plays an important role in tuning material properties and provides an opportunity to engineer a metallic alloy for a given task.

Due to the limited spatial resolution of APT, crystallographic defects like dislocations cannot be directly extracted from atomic positions in the APT reconstructed data. Only the dislocations which are decorated with solute atoms can be visualized and analyzed in 3D via APT owing to its high chemical resolution. Decorated dislocations in the reconstructed data can have linear morphology (elongated in one direction or along an arc ) or can appear as loops as in a nickel-based superalloy analyzed in [102]. APT-based analysis has improved the understanding of chemical segregation at dislocations. For example, Medouni et al. [98] explored in detail the effect of the elastic field associated with dislocation on solute segregation. Additionally, Zhou et al. analyzed the influence exerted by the structural attributes and relative orientations of dislocations on chemical segregation [97].

A typical APT analysis of dislocations comprises calculating 1D composition profiles along and perpendicular to the dislocation by manually placing primitives of, for example, cylindrical shapes [103]. Further local proxigrams are also used to quantify radial composition fluctuation from the dislocation core [98]. If dislocations are curved or form complex networks, the analysis becomes tedious and cumbersome. In a study analyzing irradiated reactor pressure vessel steels via APT, the

composition along curved dislocations was extracted by manually placing points along the dislocation core to fit a spline [104]. To make the typical analysis of dislocations more automatic, Felfer et al. relied upon the distance from the center of mass (DCOM) approach to capture individual line-like features from reconstructed data [103]. Voronoi tessellation is then used for further analysis of dislocations in their study. However, their method lacks the flexibility to analyze dislocation networks or more complicated morphologies that are often found in reconstruction data.

Circumventing the above issues, Ghamarian et al. proposed a scheme based on skeletonization [105], [106]. Skeletons are powerful abstractions for a 3D object in the form of a 1D graph [107]. They conserve the topology and geometry of the underlying object. In this approach, the skeleton of a microstructural feature in the reconstruction data is extracted using the positions of the atoms lying in the feature. The skeleton morphology then helps to find line-like features in the data. Based on the skeleton, further analysis, like the calculation of radial proxigrams was also devised. In the workflow, the atoms pertaining to a microstructural feature are obtained by applying a clustering algorithm on the spatial coordinates of the atoms, which is computationally demanding considering the large number of atoms in a typical APT data set. Furthermore, the skeletonization scheme used in their study is based on the algorithm proposed by Cao et al. in [106], which does not guarantee the alignment of the calculated skeleton with the medial axis of the 3D object. Alignment with the medial axis is necessary to compute accurate composition profiles for a dislocation. Although their method compares the morphology of different microstructural features to extract line-like features, it does not elucidate how to segment dislocations if they are protruding out of a blob-like morphology, for example, from another phase. The approach also does not take into account the local morphology of dislocation; for example, what appears to be a single line feature in APT can be two partial dislocations separated by a stacking fault.

In this chapter, an automatic and general workflow is devised to extract and analyze line-like (or linear) features in APT reconstruction data. Since iso-concentration surfaces (or isosurfaces) are faster to compute compared to clustering, they are used to isolate various microstructural features. The medially centered skeletons of the isosurface meshes are calculated. I will show how this description of skeletons can be used to segment out each dislocation as a separate entity, even in microstructures with complex morphology, by application of the workflow on both synthetic and experimental data. The skeletonization forms the basis for placing geometric primitives or regions of interest (ROIs) on the segmented dislocations in an automatic way while considering their curvature for robust composition analysis. Furthermore, the local morphology of the segmented dislocation meshes is used to define the ROIs for extracting composition profiles that are most meaningful from the materials science point of view. Finally, the orientation of the segmented dislocation in the crystal coordinate system is calculated using crystallographic information obtained from the APT data.

## 5.2 Materials and Methods

### 5.2.1 Experimental Data

The APT reconstruction data corresponding to a Ni-based single-crystal superalloy is used to showcase the segmentation of the dislocations based on the local morphology of the isosurfaces. The composition of the alloy in weight percent is given in Table 5.1. The experimental details for this alloy corresponding to APT measurements and sample preparation are mentioned in [108].

Further, reconstruction data for a Fe-Mn alloy is used to demonstrate dislocation orientation analysis in APT. The composition of the alloy in weight percent is given in Table 5.2. The details related to APT measurements and sample preparation for this alloy are discussed in [109].

TABLE 5.1: Composition of Ni-based single-crystal superalloy in wt.%. The rest of the amount is Ni.

| Cr | Co | W | Al | Ti | Nb | Hf | B | Fe | C | Zr |
|---|---|---|---|---|---|---|---|---|---|---|
| 8.60 | 9.50 | 11.80 | 4.9 | 1.87 | 0.86 | 1.58 | 0.015 | 0.02 | 0.013 | 0.01 |

TABLE 5.2: Composition of Fe-Mn alloy in wt.%. The rest of the amount is Fe.

| Mn | C | Ni | Co | Mo | Si | Al | S | P | O | N |
|---|---|---|---|---|---|---|---|---|---|---|
| 8.46 | 0.0075 | 0.0175 | 0.0022 | $< 0.002$ | 0.0024 | $< 0.002$ | 0.0047 | $< 0.002$ | 0.0102 | 0.0040 |

### 5.2.2 Workflow

Figure 5.1 summarises the workflow developed to systematically segment and quantify dislocations in APT data. As in previous chapters, AP Suite is used to reconstruct the APT data to obtain the spatial coordinates (x,y,z) and the correspondence between the mass-to-charge state of each ion and its elemental or molecular identity. Subsequently, isosurface meshes are extracted using the marching squares algorithm implemented in Paraprobe [82] for a given isovalue. This is followed by extraction of the mean curvature flow skeletons using CGAL (Computational Geometry Algorithms Library), which is an open-source C++ based computational geometry library, as explained in section **5.2.3** [110]. If the isosurface meshes consist of a complex morphology or a network of entangled linear features, a segmentation based on connectivity is performed on the obtained skeleton; refer to section 5.2.5. A B-spline is fitted onto each skeleton segment. The fitted B-splines provide the basis for a robust and automatic description of the cylindrical and radial regions of interest (ROIs), therefore allowing a high throughput analysis, section 5.3.1. The curve skeletonization captures the topology of the underlying mesh and thus allows morphology, composition, and orientation-based analysis of the linear features. After the analyses, all the corresponding descriptors are systematically stored in hdf5 files [88].

### 5.2.3 Curve skeleton extraction

Curve skeletons are widely used 1D constructs that encapsulate the geometry and topology of a 3D mesh or point cloud in a simplified manner [106], [107], [111]. Here, we extract curve skeletons from the 3D meshes pertaining to isosurfaces using mean curvature flow (MCF). We have used CGAL's implementation, based on the approach developed by Tagliasacchi et al., to calculate MCF skeletons [107], [110]. Tagliasacchi et al.'s [107] formulation and algorithm are discussed below.

A watertight surface $F$ can be approximated by a triangular mesh $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ contains vertices $[v_1, v_2, ..., v_n]$ and $\mathbf{E}$ contains edges $e_{ij}$ joining vertices $i$ and $j$. The MCF algorithm moves the points on surface S iteratively in the anti-normal direction with a speed proportional to the mean curvature at each point, as shown in equation 5.1 [107].

$$\dot{F} = -H\mathbf{n} \tag{5.1}$$

, where

$$H = \frac{k1 + k2}{2} \tag{5.2}$$

In equation 5.2, $H$ is the local mean curvature and $k1$ and $k2$ are the local principal curvatures. MCF is an iterative surface smoothening or fairing algorithm. It iteratively reduces the surface area of a mesh. Since a curve skeleton has volume and surface area equal to zero the MCF algorithm drives the mesh towards such a manifold by a reduction in surface area. To solve MCF on a discretized grid

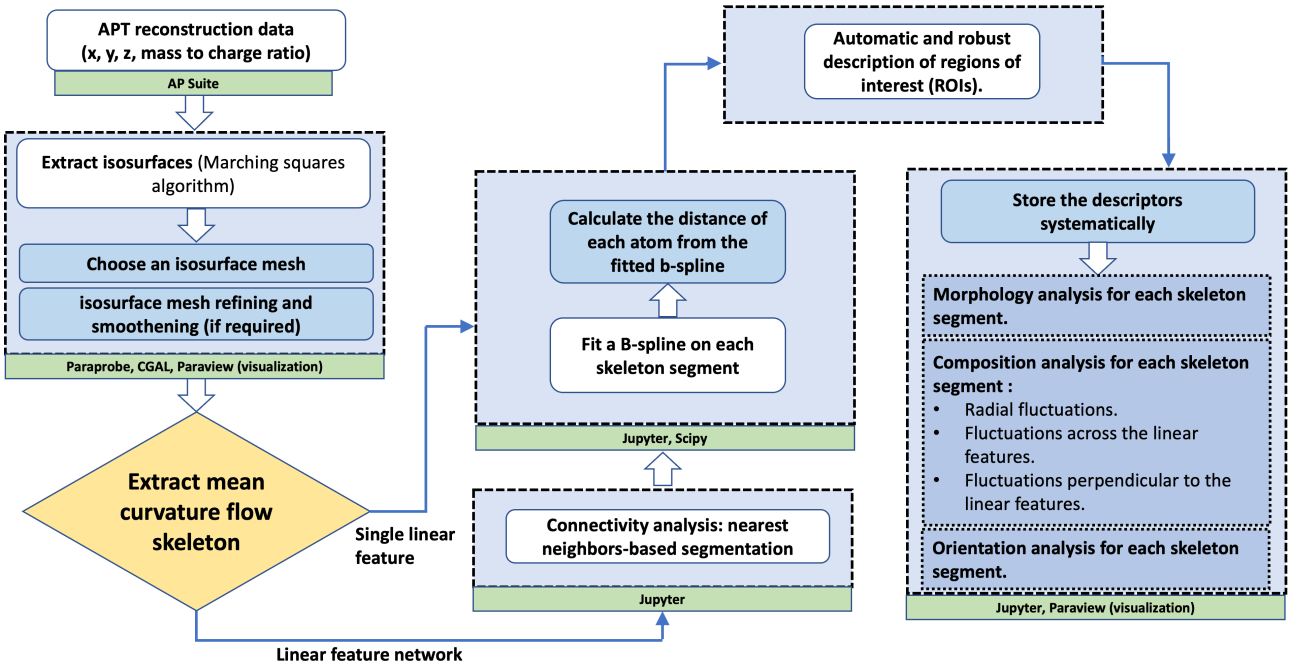



FIGURE 5.1: The workflow to extract and quantify linear features from APT reconstruction data. The tools and libraries used are mentioned in the green boxes.

implicitly, the Laplacian operator, $\mathbf{L}$, as formulated in [112], is used. Since the Laplacian operator is singular, solving equation 5.3

$$\mathbf{LV} = 0 \tag{5.3}$$

leads to a trivial solution. Circumventing this issue the following extended form is solved instead

$$\begin{bmatrix} \mathbf{W}_L\mathbf{L} \\ \mathbf{W}_H \end{bmatrix} \mathbf{V}^{t+1} = \begin{bmatrix} 0 \\ \mathbf{W}_H\mathbf{V}^t \end{bmatrix} \tag{5.4}$$

where $\mathbf{W}_L$ and $\mathbf{W}_H$ are diagonal matrices explained below [107]. The linear system of equations in 5.4 can be solved by minimizing the energy E.

$$E = \|\mathbf{W}_L\mathbf{L}\mathbf{V}^{t+1}\|^2 + \Sigma_i \mathbf{W}_{H,i}^2 \|\mathbf{v}_i^{t+1} - \mathbf{v}_i^t\|^2 \tag{5.5}$$

The first part of equation 5.5 is called $E_{smooth}$ and it contracts the given mesh. The second part corresponds to $E_{velocity}$ and attracts the mesh vertices to their original positions. The contraction and attraction effects are controlled by the weighing diagonal matrices $\mathbf{W}_L$ and $\mathbf{W}_H$ respectively [107][111]. Tagliasacchi et al. added another contribution to energy E called $E_{medial}$, shown in equation 5.6, that pulls the vertices of the contracting mesh towards their corresponding Voronoi medial poles. In equation 5.6, $\mathbf{W}_M$ is the weighing diagonal matrix that controls the attraction of the vertices to the Voronoi poles and $\mu$ maps the correspondence between the vertices and Voronoi poles.

$$E_{medial} = \Sigma_i \mathbf{W}_{M,i}^2 \|\mathbf{v}_i^{t+1} - \mu(\mathbf{v}_i^t)\|^2 \tag{5.6}$$

After adding all the contributions, the final energy becomes,

$$E = E_{smooth} + E_{velocity} + E_{medial} \tag{5.7}$$

The pseudocode for the calculation of the MCF skeleton is shown in Algorithm 1. A triangular mesh corresponding to a water-tight surface $F_{in}$ is input into the MCF skeleton algorithm. First, as a preprocessing step, the Voronoi diagram of the given mesh is extracted to get the medial poles. The Voronoi medial poles lie on the medial axis of the given mesh. Next, the MCF contraction is applied to the mesh iteratively. After every iteration since the given mesh contracts, therefore curvature at each vertex changes. These changes are accommodated by updating the Laplacian operator and weights. Also, after each contraction, some triangles can be badly shaped, or some edges might be reduced to very small lengths. To remove these, a re-meshing step is performed. Finally, after getting a manifold with zero volume, all the short edges are collapsed to achieve a curve skeleton [107]. In CGAL's implementation of MCF skeleton algorithm, the initial values of $\mathbf{W}_H$ and $\mathbf{W}_M$ have to be provided by the user.

Input: Triangulated surface $F_{in}$
Output: Mesh curve skeleton
**function** Calculate mesh skeleton(Input: $F_{in}$)
  F← Surface mesh re-sampling(Input: $F_{in}$)
  pol← Compute voronoi poles of the mesh(Input: $F$)
  **while** *still has volume* ($F'$) **do**
    Laplacian and weights are modified(Input: $F$, pol)
    Mesh contraction is applied(Input: $F$, pol)
    Mesh connectivity is edited(Input: $F$)
  **end**
  Shortest edges of mesh triangles are collapsed(Input: $F$)
  **return** $F$
**end function**

**Algorithm 1:** Pseudocode for curve skeleton calculation (reproduced from [107])

### 5.2.4 Distance from a B-spline

A cubic B-Spline is fitted on the skeleton-segment nodes that delineate a linear feature, shown in Figure 5.2. The fitted spline, $J(u)$, is parameterized by $u(0 \leq u \leq 1)$ and is defined by the knots vector $k$ and coefficients $c$. As the linear features in the isosurfaces can be dislocations, it is important to quantify which chemical species have a higher affinity towards them for material design. To quantify the affinity towards the dislocation, the minimum distance of each atom is calculated from its geometric core or the fitted B-spline. As the knots and the coefficients of the spline are known, the distance$(D)$ between any atom $P'(X', Y', Z')$ and a point, $P(x(u), y(u), z(u))$, on spline for a given value of $u$ is given in equation 5.8. A simplex optimization scheme is performed on equation 5.8 to find the value of the parameter $u$ corresponding to the minimum distance ($d$) between point $P'$ and the fitted B-spline [113] [114]. Since the optimization problem for each atom is independent, the algorithm has been parallelized to make the computation of distances for millions of atoms faster and scalable.

$$D = \sqrt{(X' - x(u))^2 + (Y' - y(u))^2 + (Z' - z(u))^2} \tag{5.8}$$

### 5.2.5 Validation on synthetic data

To showcase the main features of the developed workflow (Figure 5.1), it has been applied to two synthetic datasets. An artificial data set is created with two $L1_2$-ordered structure ($Al_3$-Li) cylinders

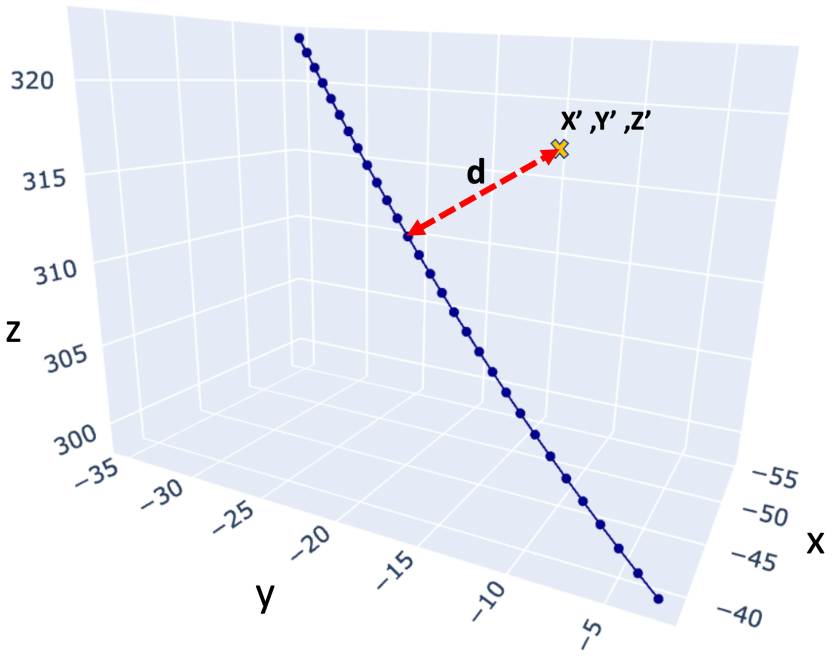



FIGURE 5.2: A B-spline fitted on skeleton nodes is shown. Here, d is the minimum distance of an atom with coordinates (X', Y', Z') from the B-spline.

($r_{in}$ = 1.5nm) crossing each other at 90° in a pure Al matrix, shown in Figure 5.3(A). The dataset corresponding to this morphology is referred to as cross-dataset. The second synthetic dataset contains a half torus ($r_{in}$ = 3nm and $r_{out}$ = 20nm) of $L1_2$ structure in an Al matrix, Figure 5.4(A). For both datasets, isosurfaces for 11at% Li are calculated.

The cross-dataset demonstrates the segmentation of the calculated skeleton as shown in Figure 5.3(B). The skeleton for the 11at% Li isosurface mesh is computed. A connectivity analysis is performed on the obtained 1D skeleton. The skeleton nodes that have more than two connected neighbors are considered as junction nodes and the ones with just one connected neighbor are the end nodes, Figure 5.3(B). A skeleton segment is then defined as the set of nodes that start from an end node and end at a junction node, start at a junction node and end at another junction node or start at an end node and end at another end node. Having applied the segmentation analysis on the cross-dataset, four segments are obtained, as shown in Figure 5.3(C). The segmentation scheme exemplified here can be used to segment and analyze entangled linear features in APT data.

The application of the MCF skeletonization algorithm to obtain the curve skeletons conserves the topological information of the isosurface mesh and also makes sure that the skeleton is close to the medial axis of the input mesh. The MCF skeletonization is applied on the 11at% Li isosurface mesh in the half-torus dataset to get the medial skeleton as shown in Figure 5.4(B). Since the geometric medial axis of the half torus is already known, this synthetic data helps to measure the accuracy of the MCF algorithm on curved linear features. A thorough analysis of the algorithm's performance for this dataset is presented in the section 5.3.1. Further, a B-spline is fitted on the skeleton, and each atom's shortest distance is calculated from the fitted spline, as shown in Figure 5.4(C).

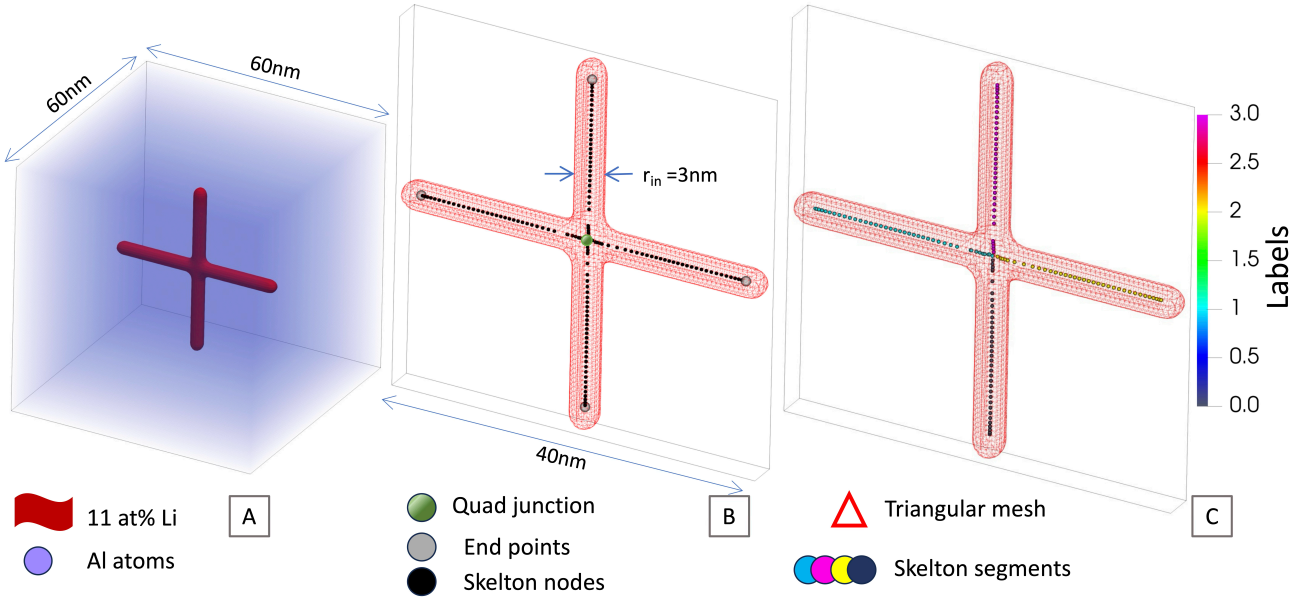



FIGURE 5.3: (A) 11at.% Li isosurface mesh with cross morphology is computed. (B) Mean curvature flow skeleton (MCF) for the mesh is shown with the junction and end nodes highlighted. (C) Each branch of the skeleton is segmented by connectivity analysis.

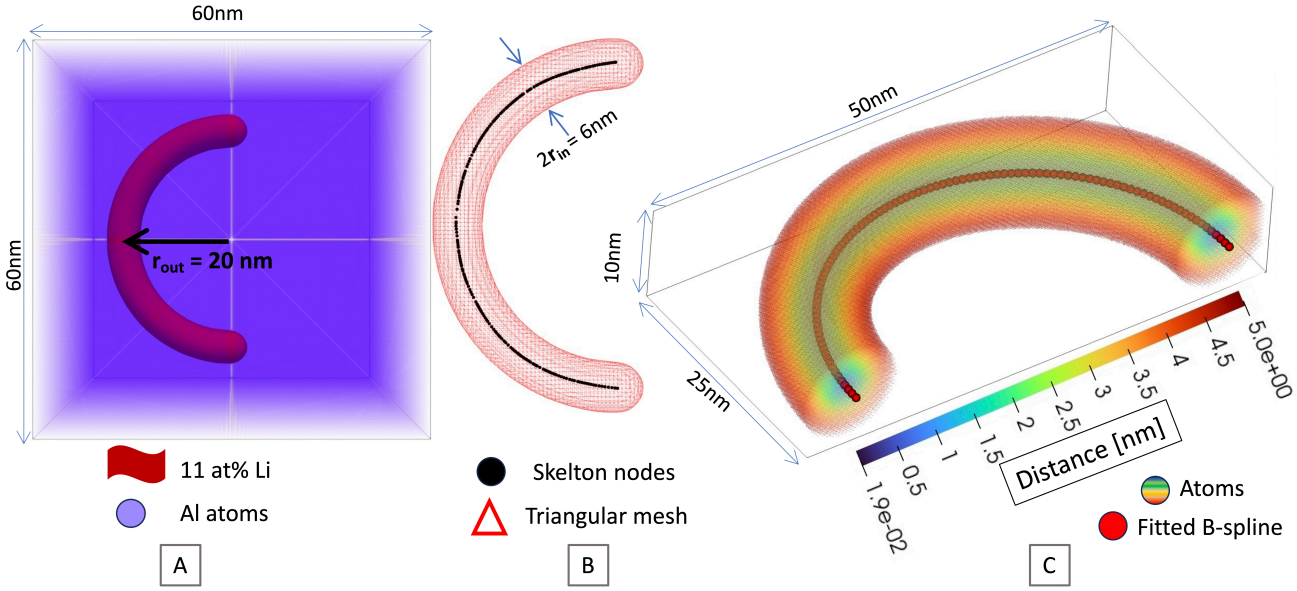



FIGURE 5.4: (A) 11at.% Li isosurface mesh with half-torus morphology is calculated. (B) The MCF skeleton for the mesh is shown. (C) Minimum distance of the surrounding atoms from the B-Spline fitted on the skeleton nodes.

## 5.3 Results and discussion

### 5.3.1 Synthetic data

I have used the half-torus dataset to analyze the performance and the effect of hyperparameter values on the results of the MCF algorithm when applied to curved meshes. As mentioned in the previous section, the isosurface mesh for the half-torus dataset is calculated. Since the actual medial axis of the half-torus is known, I calculate the average distance of the skeleton nodes obtained from the MCF algorithm from the actual medial axis as the error.

As previously stated in section 5.2.3, the initial values of $\mathbf{W}_H$ and $\mathbf{W}_M$ in the CGAL's implementation of the MCF algorithm have to be set by the user. A lower value of $\mathbf{W}_H$ leads to faster convergence of the mesh to a curve skeleton; however, the resulting skeleton may not be able to retain the intricacies of the input mesh [110]. $\mathbf{W}_M$, on the other hand, controls the attraction of the nodes of the intermediate meshes (or mesoskeletons) towards the medial axis of the input mesh. The impact of the initial values of $\mathbf{W}_H$ and the resulting skeletons for the half-torus isosurfaces ($r_{out} = 20$nm and $r_{out} = 10$nm) is elucidated in Figure 5.5 (A and B). For both half-torus meshes the error or the deviation from the medial axis drops with an increase in the value of $\mathbf{W}_H$. However, the rate of drop in error is higher in the mesh with lower curvature ($r_{out} = 20$nm). So, for linear features in APT data with low curvature, which is frequently the case, even low values of $\mathbf{W}_H$ will lead to correct skeletons with faster convergence. After $\mathbf{W}_H = 0.05$ the error drops to approximately 0.01nm for both the meshes. The distribution of distances from the true medial axis for all the skeleton nodes for the mesh with $r_{out} = 20$nm and $\mathbf{W}_H = 0.05$ is illustrated in Figure 5.6. Although the error of 3.5 nm, corresponding to $\mathbf{W}_H = 0.004$ for the mesh with $r_{out} = 10$nm, might seem insignificant from the perspective of the relevant length scales in APT data but as shown in Figure 5.5 (A) the corresponding skeleton completely fails to capture the morphology of the underlying input mesh. The effect of change in $\mathbf{W}_M$ value for a given value of $\mathbf{W}_H$ on the calculated skeleton is not significant as the change in error is of order of $10^{-3}$nm for the given meshes. However, if the medial contribution is completely switched off, the convergence to the medial axis is not guaranteed. For all the following cases the initial value of $\mathbf{W}_H = 0.05$ and $\mathbf{W}_M = 0.2$ (which is the default value of $\mathbf{W}_M$ in CGAL).

Once the curve skeleton of the half-torus ($r_{out} = 20$nm) is calculated, a B-spline is fitted on the skeleton nodes. This is followed by obtaining the closest distance of atoms from the computed spline. The radial fluctuations in composition from the fitted spline can be calculated by defining bins as shown in Figure 5.7(A). Each bin contains atoms that are within a distance range of $r'_{out}$-$r'_{in}$ (here, $= 1$nm) from the fitted spline or the geometric core of the linear feature. As shown in the figure, selecting atoms in this way ensures that the bins take the shape of a linear feature, like concentric sleeves. Further, the average composition of each bin is calculated, refer 5.7(B). As the half-torus consists of $L1_2$ structure ($Al_3Li$), it contains 25at% Li, which can also be seen in the composition vs distance (from the spline) plot in Figure 5.7(B). Note that after 3nm, the composition of Li falls to 0at% as the radius of the half-torus is 3nm, and the matrix is made of pure Al.

From a materials science point of view, if a linear feature in APT data corresponds to a dislocation, the quantification of composition fluctuations along and perpendicular to the linear features is of immense importance. Composition fluctuations along the dislocation may be correlated to secondary phase formation or spinodal decomposition [115]. Perpendicular fluctuations can help estimate the affinity of a chemical species towards the dislocation or maybe conclude that the linear feature is actually made of two partial dislocations. For this, a robust and automatic description of ROIs is developed, as shown in Figure 5.8.

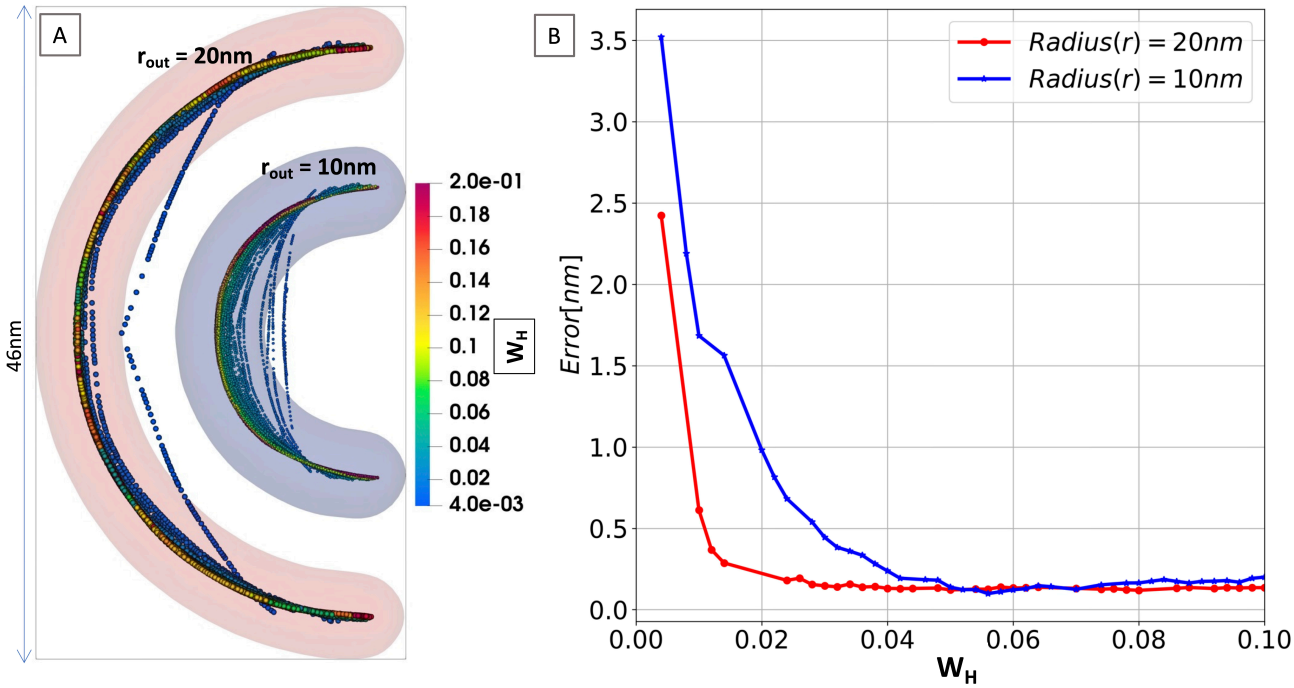



FIGURE 5.5: (A) MCF skeletons for half-torus meshes with different curvatures ($r_{out}$ = 20nm and 10nm) and varying value of $W_H$ hyper-parameter. (B) The variation of error or the average deviation from the true medial axis of the torus, for each computed skeleton, with values of $W_H$ is plotted for both the meshes.

The fitted B-spline is used as the basis for describing the ROIs. To capture the composition fluctuations along the dislocation, first, all the atoms within a distance $r$ (here, = 2.5nm) from the B-spline are selected. Since the B-spline is parametrized by $u$ ($0 \leq u \leq 1$), ROIs are described by dividing $u$ into equal intervals and selecting atoms in each interval. This description ensures that each ROI takes into account the local curvature of the dislocation and there is no overlap between the ROIs as shown in Figure 5.8(A). The corresponding 1D composition profile, calculated by taking the average composition of each ROI, is shown in Figure 5.8(B). A cylindrical ROI, which is a more typical description, along the dislocation can be described by choosing the vector, $\vec{k}$, joining two B-spline nodes as the axis, Figure 5.8(C). The corresponding 1D composition profile is shown in Figure 5.8(D). Longer cylindrical ROIs may not be able to capture the curvature of the dislocation correctly and lead to errors in composition fluctuations along the dislocation. Also, with an increase in curvature, the overlap between ROIs increases. However, here, as the ROI lengths are shorter, the cylindrical description yields similar results as the first description with the difference in average composition to be only 0.14 at.% Li. Cylindrical ROIs perpendicular to the spline at a given spline node can be defined by finding vectors perpendicular to $\vec{k}$ and making an angle $\theta$ with the z-axis as shown in Figure 5.8(E). The corresponding 1D composition profile along one such ROI shows composition across the half-torus to be ~25 at.% Li, Figure 5.8(F).

### 5.3.2 Morphology based segmentation

In APT data, dislocations where segregation has happened appear as linear features. However, there can be cases where dislocations, due to prior processing, penetrate from one phase to another. For example, in nickel-based superalloys, dislocations enriched in Cr, Co, and Fe extend into $\gamma'$ phase from $\gamma$ phase [116]. In these cases, the linear features appear to be part of a larger non-pipe-like feature, Figure 5.9(A). Also, as dislocations can be preferential nucleation sites, there can be cases as

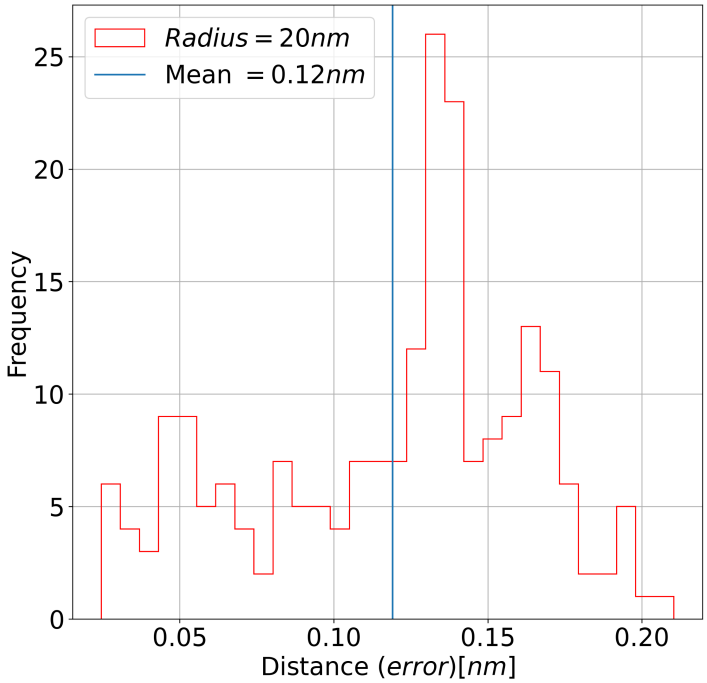



FIGURE 5.6: The distribution of the distances from the true medial axis for all the skeleton nodes for the half-torus mesh with $r_{out} = 20$nm and $\mathbf{W}_H = 0.05$.

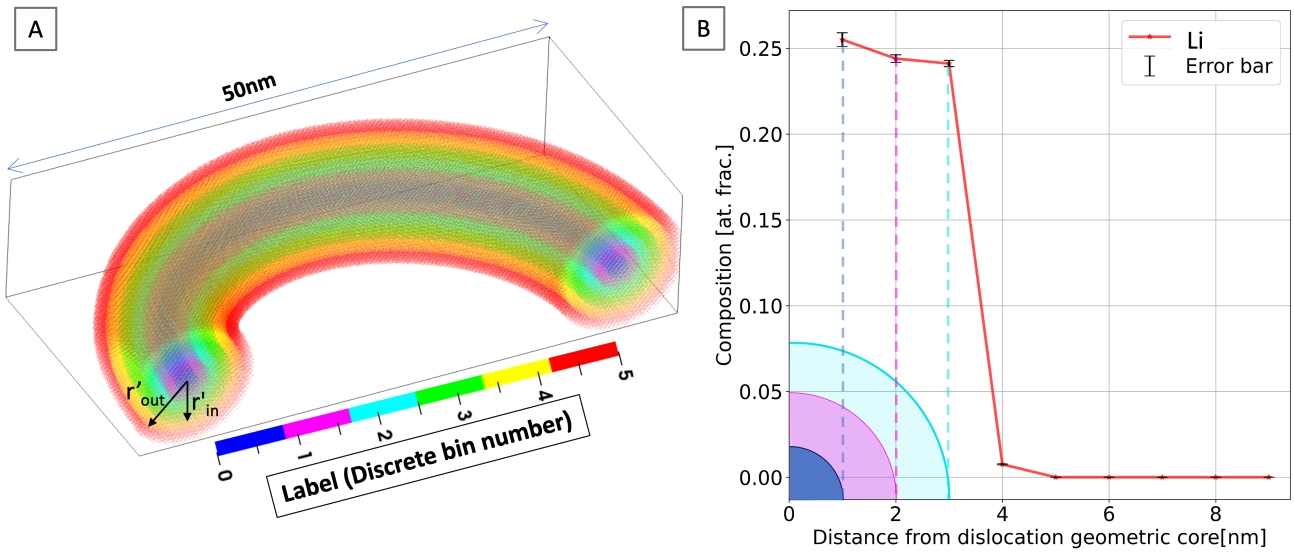



FIGURE 5.7: (A) Different colors illustrate various bins. Each bin contains atoms that are within the distance range of $r'_{out} - r'_{in} = 1$nm from the dislocation geometric core. (B) Radial fluctuation in the composition is computed. The colored quarter circles depict the corresponding bin in panel (A), whose composition is calculated. Error bars correspond to $2\sigma$ counting error.

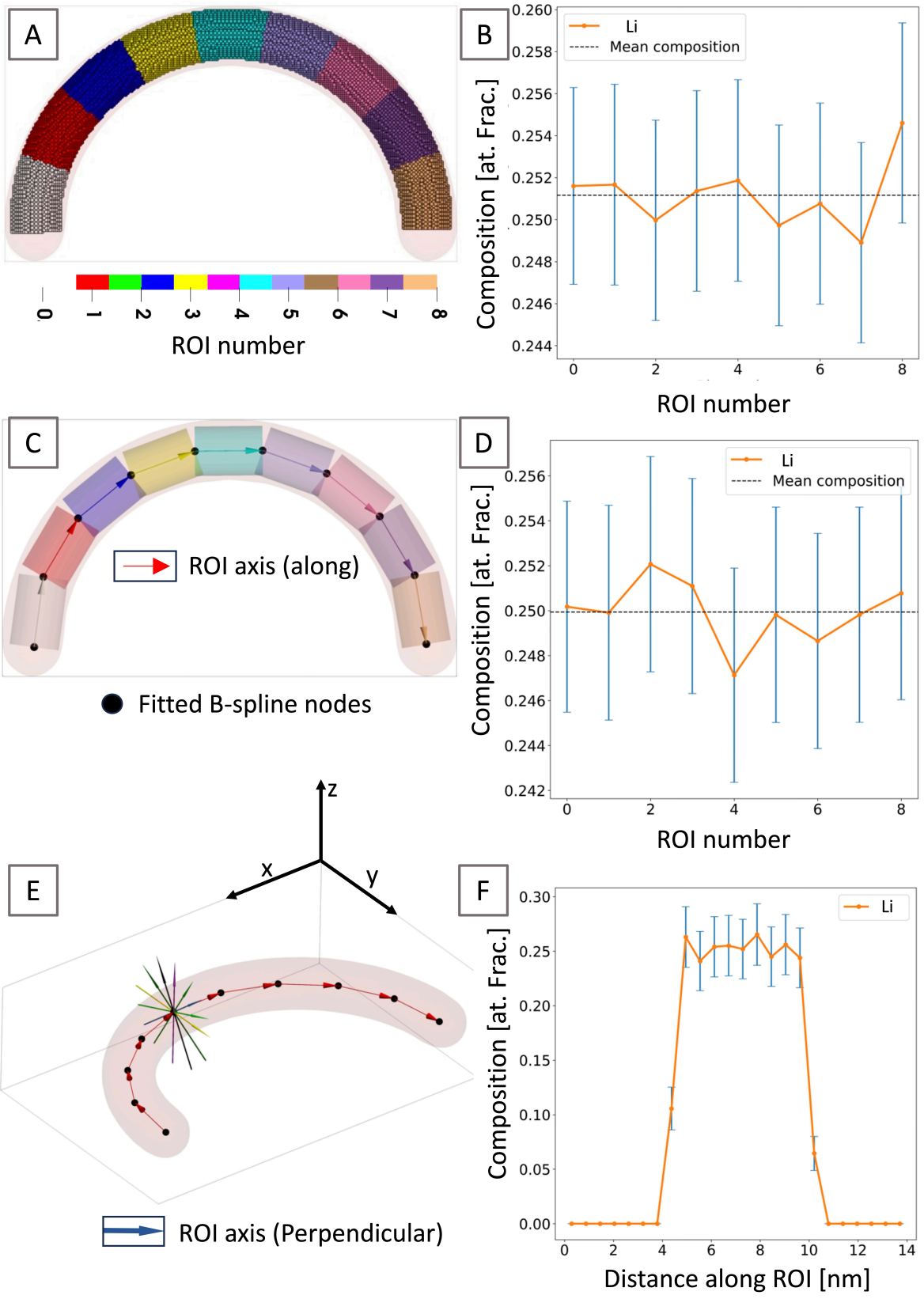



FIGURE 5.8: Different ways of placing the ROIs on a half-torus linear feature/dislocation using the fitted B-spline. (A,B) ROIs along the dislocation without any overlap with the corresponding fluctuations in composition along the linear feature. (C,D) Cylindrical ROIs with overlap and composition fluctuations along the linear feature. (E,F) The arrows define the axis for the cylindrical ROIs perpendicular to the linear feature with composition fluctuations across the linear feature for one such ROI.

in [109], where linear features appear to come out of a blob-like feature. Analyzing these complex morphologies can be important for understanding physical phenomena like pipe diffusion [116] and nucleation.

Figure 5.9(A) shows $6at.\%$ Cr isosurface in a Ni-based single crystal superalloy [108]. A dislocation network extends from the $\gamma$ matrix into a $\gamma'$ precipitate. MCF curve skeleton nodes for the isosurface mesh are shown in Figure 5.9(B). The nodes identified as ends and junctions are also marked in the figure. Using the segmentation scheme based on the connectivity analysis in section 5.2.5, 74 segments are recognized, shown in 5.9(C). Clearly, the current scheme (primary segmentation) fails to automatically distinguish skeleton segments corresponding to the dislocation network from other segments. This is often the case when the linear features originate from larger non-pipe features or blobs.

Isolating the linear features automatically is necessary for robust and high-throughput analysis. Since the mesh vertices are contracted iteratively to the skeleton nodes, each node can be traced to its parent vertices. So after the primary segmentation of the skeleton, each segment, which consists of nodes, is traced back to its corresponding mesh vertices as shown in Figure 5.10(A). The primary segmentation of the skeleton forms the basis for the triangular mesh segmentation. This is followed by calculating the PCA vectors and their explained variance ratios for the set of each segmented mesh vertices. The distribution of PC1 explained variance ratio ($PC1_{ratio}$) for all the segmented meshes is shown in Figure 5.10(B). The meshes with higher PC1 explained variance contribution ($PC1_{ratio}$) are expected to be elongated and may have a linear morphology. All meshes with $PC1_{ratio} \geq 0.8$ are filtered out as possible linear features, 5.10(C). As depicted in the figure, the dislocation network is still not selectively isolated. This is because the meshes with a non-tube-like morphology may still have PC1 considerably larger than the other two components (PC2 and PC3); hence $PC1_{ratio}$ will be larger and can be confused with tube-like features using the above criteria.

As shown in Figure 5.10(D), the final filtering step is performed on the selected meshes. For instance, for the selected meshes 1 and 2 in the figure, PCA vectors for the mesh vertices and corresponding underlying skeleton segment nodes are calculated. If the angle ($\theta$) between the PC1 vectors of the skeleton nodes (PC1(Skel)) and mesh vertices(PC1(Ver)) is approximately zero, that would mean that the mesh segment is a linear feature. The distribution of $Cos(\theta)$ is shown in 5.10(E). All the meshes with $Cos(\theta) > 0.99$ are the features with linear morphology as marked by a circle in Figure 5.10(C).

Using the skeleton segment for the dislocation labeled as 2 in 5.10(D), atomic distances can be calculated from the geometric core of the dislocation as discussed in section 5.3.1. The distances are used to compute the composition fluctuations with distance from the dislocation core for various chemical species, as illustrated in Figure 5.11. Evidently, Cr and Co have a higher tendency to segregate at the dislocation compared to the $\gamma'$ precipitate. Further, the composition of Cr falls off faster than that of Co as the distance from the dislocation core increases, suggesting a higher affinity of Cr towards the dislocation core relative to Co.

Additionally, the segmented mesh vertices pertaining to the above dislocation can be utilized to gain more information about the dislocation morphology. PC1(Ver) vector for this segmented mesh aligns with its length. However, PC2 and PC3 vectors for the mesh vertices will orient themselves perpendicular to the length of the dislocation and along its cross-section, as illustrated in Figure 5.12(A). The relative explained variance of the PC2 and PC3 conveys if the cross-section is circular or elliptical. Figure 5.12(A) shows that PC3 vector is perpendicular to the surface of the mesh with lower curvature, and PC2 is perpendicular to the surface with higher curvature. Moreover, vectors parallel to the PC2 and PC3 are used as axes for cylindrical ROIs placed automatically along the fitted B-spline

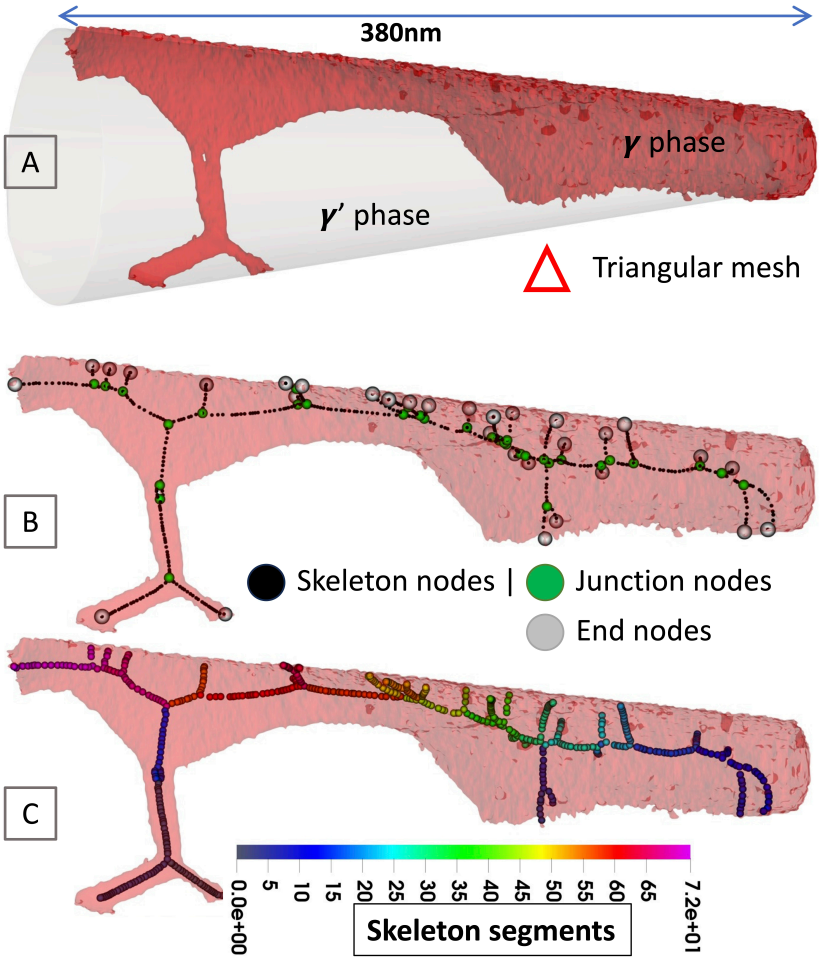



FIGURE 5.9: (A) 6at.% Cr isosurface mesh (Red). A dislocation network protrudes from $\gamma$ phase to $\gamma'$ precipitate. (B) MCF skeleton with end and junction nodes for the mesh. (C) Connectivity-based segmentation is performed on the computed skeleton. Skeleton segments are shown in different colors.

on the skeleton nodes of the dislocation, shown in Figure 5.12(B). 1D composition profiles for Cr along the ROIs parallel to PC3 and PC2 are shown in Figure 5.12(C) and (D) respectively. The two peaks in the 1D composition profile for Cr (Figure 5.12(D)) along ROI parallel to PC2 show the possibility of the existence of a staking fault. Furthermore, ROIs along the dislocation, considering its curvature, are defined using the scheme described in section 5.2.5, are shown in Figure 5.12(E). The corresponding 1D composition profile for Cr is shown in panel (F) of the figure. Therefore, this method enables an automatic composition quantification of linear features with complex morphologies.

### 5.3.3 Orientation analysis

In this section, the scheme to extract the orientation of the dislocations in APT data relative to the crystal coordinate system is explained and discussed. The crystallographic information can be retrieved from the APT data using the low-density regions called poles that form on the detector hit map [7], [31]. Although the region around a pole has low atomic density, the spatial resolution in the depth direction is better than other regions [7], [31]. The angular relationships and spatial orientation of poles on the hit map serve as indicators for identifying corresponding crystallographic planes [7], [41]. For an accurate 3D reconstruction of the APT data, the angle between the poles and the distance

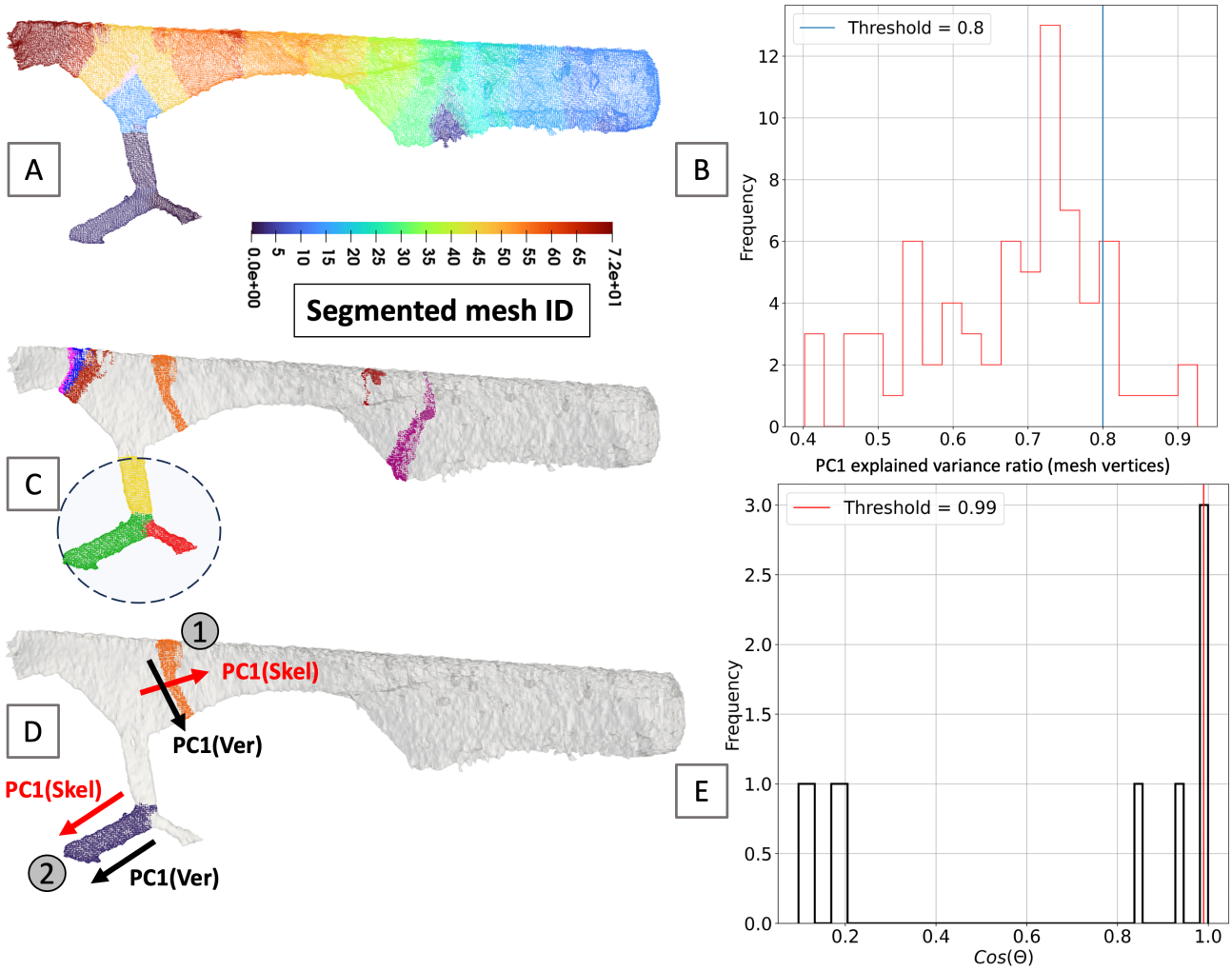



FIGURE 5.10: (A) Segmented mesh vertices corresponding to skeleton segments in Figure 5.9 (C). (B) The distribution of PC1 explained variance ratio ($PC1_{ratio}$) calculated for the set of each segmented mesh vertices is shown. (C) Segmented meshes with $PC1_{ratio} \geq 0.8$. The meshes forming the dislocation network are circled. (D) The PC1 vectors for skeleton nodes(PC1(Skel)) and mesh vertices (PC1(Ver)) are shown for two segmented meshes. (D) Distribution of $Cos(\theta)$ where $\theta$ is the angle between PC1(Skel) and PC1(Ver) vectors. Meshes with $Cos(\theta) \geq 0.99$ are considered to have pipe-like or linear morphology.

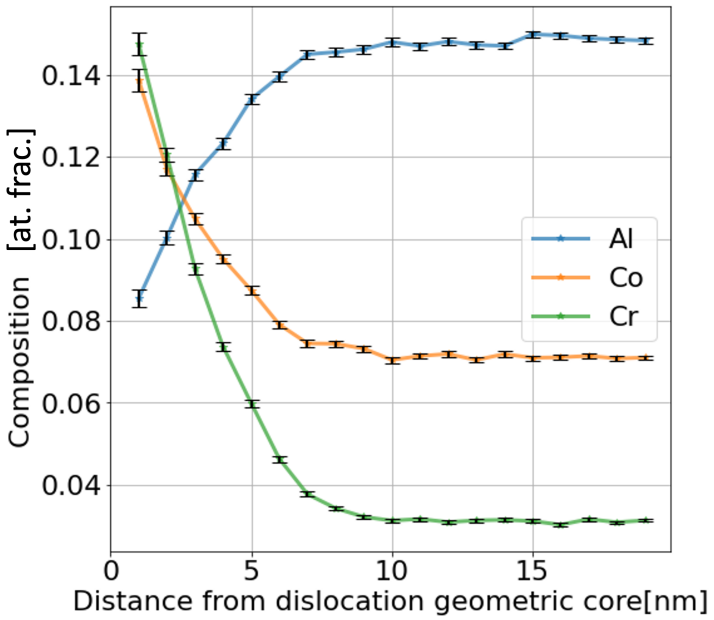



FIGURE 5.11: Radial fluctuations in the composition of Al, Co, and Cr about the geometric core of the dislocation marked 2 in the Figure 5.10(D).

between the crystallographic planes at a given pole help to fix the reconstruction hyper-parameters: image compression factor and field factor ($k_f$) [79].

To have an optimum reconstruction of the APT data, it is reconstructed at a pole such that the z-axis of the reconstruction coincides with the pole. This also makes sure that the corresponding crystallographic planes, Miller indices (hkl), of the pole are perpendicular to the z-axis. For a cubic system, the z-axis will lie along the [hkl] direction in the crystal coordinate system.

The challenge is that to completely transform the reconstructed APT data coordinate system to crystal coordinate system, Miller indices of at least 2 vectors must be known. Therefore, apart from the pole that coincides with the z-axis, crystallographic planes of other poles and corresponding directions have to be identified. Consequently, I have modified the algorithm of Araullo-Peters et al. to find crystallographic planes and normals in reconstructed APT data automatically [117]. The modified method is explained below.

#### Orientation analysis algorithm

Atoms of a particular chemical species in a spherical ROI of radius 2 nm are considered around a pole, as shown in Figure 5.13 (A). Next, the origin is shifted to the centroid of the atoms in the ROI. Further, a structured grid is constructed comprising ordered pairs of azimuthal ($\theta$) and zenith ($\phi$) angles. This grid has a resolution of 1°, and spans the angular range $0° \leq \theta \leq 360°$ for the azimuthal angle and $0° \leq \phi \leq 45°$ for the zenith angle. For each ordered pair ($\theta$, $\phi$), a unit vector $\mathbf{K}(\theta, \phi)$ is formulated using the expression in equation 5.9, shown in Figure 5.13 (A). This step is followed by rotating the vector $\mathbf{K}$, with all the atoms, such that it aligns with the y-axis using the rotation matrix given in equation 5.10. Subsequently, the histogram of the y coordinates of the atoms is computed, and a 1D fast Fourier transform (FFT) is performed on the histogram. The maximum peak strength ($P(\theta, \phi)$) of the signal from the absolute values retrieved after FFT analysis is recorded after ignoring the peak corresponding to noise. The algorithm until here is repeated for each pair of $\theta$ and $\phi$. If $\mathbf{K}(\theta, \phi)$ is normal to the atomic planes, then this will be reflected in the y-coordinate histogram in the

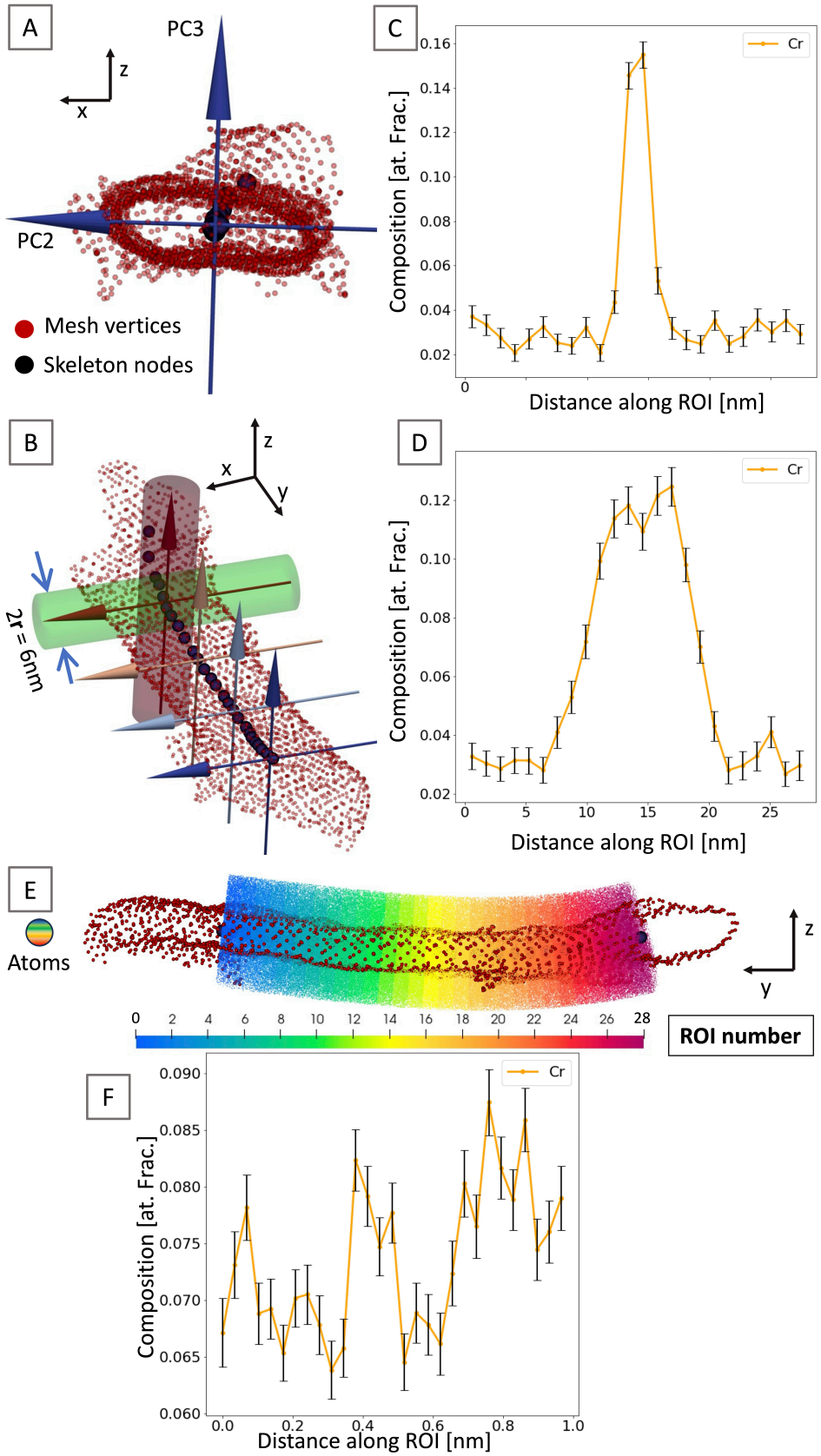



FIGURE 5.12: (A) Cross section of the mesh corresponding to the dislocation marked 2 in the Figure 5.10(D). PC2 and PC3 vectors for the mesh vertices are also shown. (B) PC2 and PC3 vectors are translated along the skeleton nodes to define cylindrical ROIs perpendicular to the dislocation. (C, D) 1D composition profiles for Cr along ROIs with PC3 and PC2 as axes, respectively. (E, F) ROIs along the fitted B-Spline on the skeleton nodes and the corresponding composition fluctuations along the dislocation.

form of periodic peaks and also in the FFT of the histogram. The vector $\mathbf{K}(\theta, \phi)$ corresponding to the maximum value of the P($\theta$, $\phi$) is the normal to the atomic planes.

After rotating the atoms and the vector **K** for a given value of ($\theta$, $\phi$), the 2d histogram of y and x coordinates of the atoms is shown in Figure 5.13 (B). Since in panel (A) of the figure **K** is normal to atomic planes, after rotation, the atomic planes become perpendicular to the y-axis as depicted in panel (B).

$$\boldsymbol{K}(\theta, \phi) = [\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)] \tag{5.9}$$

$$R = \begin{pmatrix} \cos(-\phi)\cos(-\theta) & \sin(-\phi)\cos(-\theta) & \sin(-\theta) \\ -\cos(-\phi)\sin(-\theta) & -\sin(-\phi)\sin(-\theta) & \cos(-\theta) \\ \sin(-\phi) & -\cos(-\phi) & 0 \end{pmatrix} \tag{5.10}$$

**Application on Fe-Mn alloy**

The algorithm explained above is applied to the reconstructed APT data of the Fe-Mn alloy mentioned in the section 5.2.1. Isosurface meshes corresponding to 14$at$.% Mn are shown with different colors in Figure 5.14(A). The meshes pertaining to the linear features, which are dislocations in this case, are segmented out by PCA-based filtering similar to the one explained in the section 5.3.2. The skeleton nodes for each dislocation are illustrated in Figure 5.14(B). The composition fluctuations around each dislocation can now be mapped automatically along and perpendicular to the geometric core using the skeleton as the basis. The composition fluctuation of Mn along the dislocation marked 1 in panel (B) of the figure is delineated in Figure 5.14(C). Clearly, the composition of Mn fluctuates between 15$at$.% to 24$at$.% showing the signs of the formation of a secondary phase via spinodal decomposition [109], [115].

A slice of 10nm width containing only Fe atoms is extracted from the reconstructed data of Fe-Mn alloy. The 2D histogram of the x and y coordinates of the Fe atoms is shown in Figure 5.15(B). In the 2D histogram, the regions of low atomic densities corresponding to (-101),(-111), and (011) poles or atomic planes are depicted in the figure. The reconstruction of the data was performed with the (-111) pole as the center, and it coincides with the z-axis of the reconstructed data. So, the [-111] direction (Miller indices) in the crystal coordinate system corresponds to the [0,0,1] direction in the sample coordinate system. However, for the other two poles, the corresponding direction in the sample coordinate system is unknown.

Considering the region around the (-101) pole, a spherical ROI containing Fe atoms is selected. Thereafter, for the given ROI, for all the ordered pairs of azimuthal and zenith angles ($\theta$, $\phi$), the above-mentioned algorithm is applied. All the peak strength values P($\theta$, $\phi$) acquired after FFT analysis are shown in the Figure 5.16. The maximum value of P($\theta$, $\phi$), also marked in the figure, corresponds to the normal vector to the atomic planes. The ordered pair ($\theta$, $\phi$) for the normal vector is equal to ( 349°, 35°) for the (-101) pole. Similarly, (119°, 32°) corresponds to the ordered pair for vector normal to the atomic planes at (011) pole.

Figure 5.17(A) illustrates the region around the (-101) pole from which the spherical ROI is selected for the crystallography analysis. The normal vector corresponding to the maximum value of the P($\theta$, $\phi$) is specified in panel (B) of the figure. In panel (B), crystallographic planes can be seen in the selected ROI and extending outwards from it, clearly suggesting that the normal vector is correctly calculated by the above algorithm. Also, Figure 5.18(A) shows the histogram of the y-coordinate of

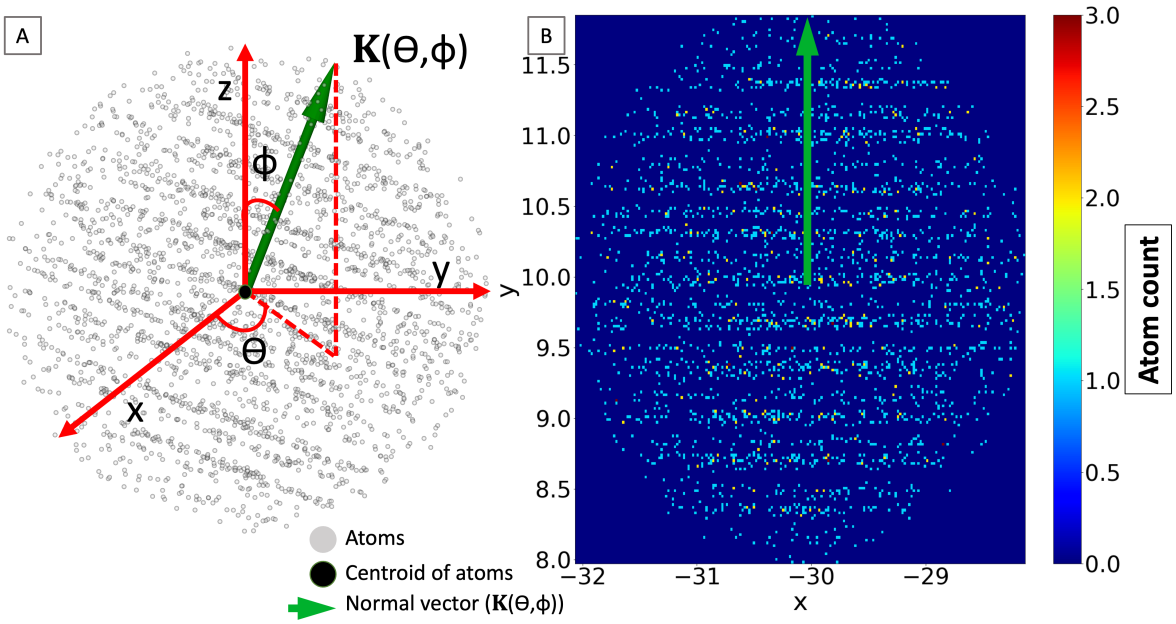



FIGURE 5.13: (A) Spherical ROI, near a pole, with atoms is shown. The vector $\mathbf{K}(\theta, \phi)$ has azimuthal angle $\theta$ and zenith angle $\phi$. (B) 2D histogram of x-y coordinates of the atoms after rotating $\mathbf{K}(\theta, \phi)$ to align with the y-axis. As $\mathbf{K}(\theta, \phi)$ is perpendicular to the atomic planes seen in panel (A) when it is rotated, the atomic planes become perpendicular to the y-axis.

the Fe atoms after the rotation of the normal vector to the y-axis. The fluctuating intensity of the peaks in the histogram also suggests the existence of atomic planes; the corresponding plot for the absolute values after FFT analysis is shown in Figure 5.18(B).

The PC1 vector of the dislocation skeleton nodes lies along the length of the dislocation. For the dislocation marked 1 in Figure 5.14(B) the PC1 vector ($PC1_{Dis(1)}$) for its skeleton nodes is calculated. The orientation of PC1 of the selected dislocation relative to the sample coordinate system (red) and to the crystal directions (green) is shown in Figure 5.19. As shown in the image, $PC1_{Dis(1)}$ makes an angle of 97° with [011] direction. This implies that the dislocation is closely aligned with the (011) plane. Given that the matrix of the alloy under consideration exhibits a body-centered cubic (bcc) structure, this finding is consistent with the well-established understanding that the (011) plane serves as a slip plane due to its closed-packed nature.

## 5.4 Conclusion

This chapter has provided a comprehensive workflow for automatically extracting and analyzing line-like features in APT reconstruction data, addressing limitations in existing approaches. The developed workflow consists of the following steps: (i) extract isosurface meshes from APT reconstruction data and filter elongated meshes based on PCA; (ii) Compute the mean curvature flow skeleton of the meshes; (iii) perform connectivity analysis based or morphology-based segmentation to extract dislocation segments; (iv) fit B-spline on each segment; (v) find distances of atoms from the spline and place ROIs on the spline for composition analysis. Apart from the composition analysis of the dislocations, crystallographic information is obtained from the APT data to calculate the orientation of the dislocation in the crystal coordinate system.

While the workflow is designed to be robust and automatic, it is essential to note that the method relies on the initial selection of the isosurface meshes, which depends on the chosen isovalue. Currently,

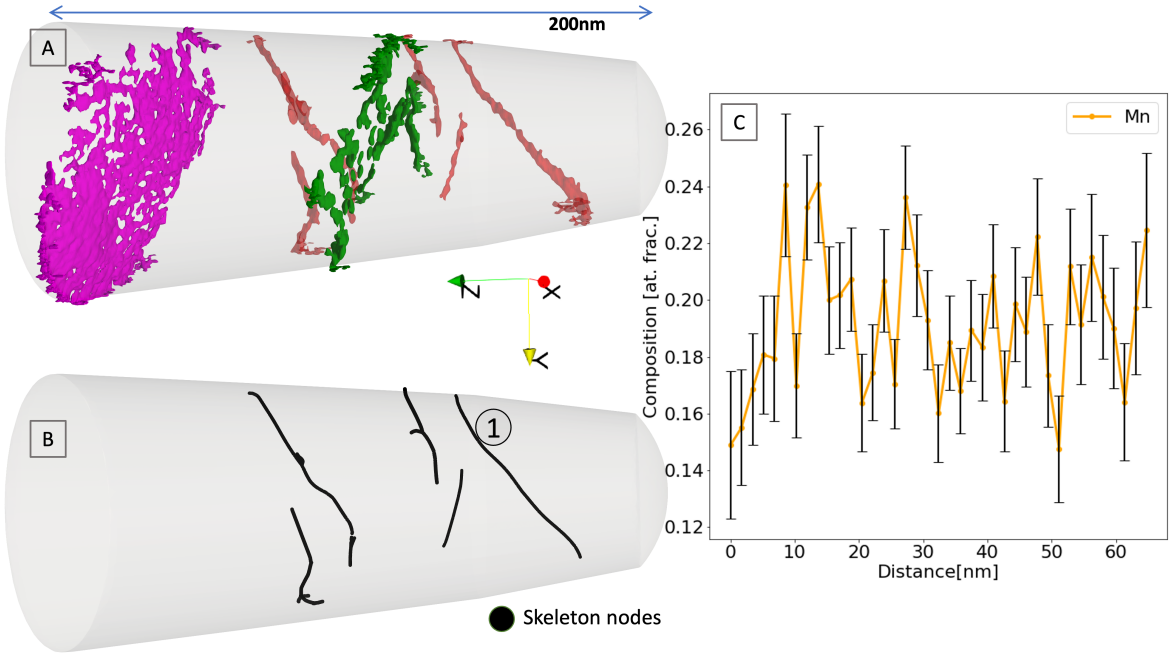



FIGURE 5.14: (A) 15at.% Mn isosurface meshes are depicted with different colors. Green, red, and pink meshes correspond to a low-angle grain boundary, dislocations, and a grain boundary. (B) MCF skeletons of all the dislocation meshes. (C) Composition fluctuations of Mn along the dislocation marked 1 in panel (B).

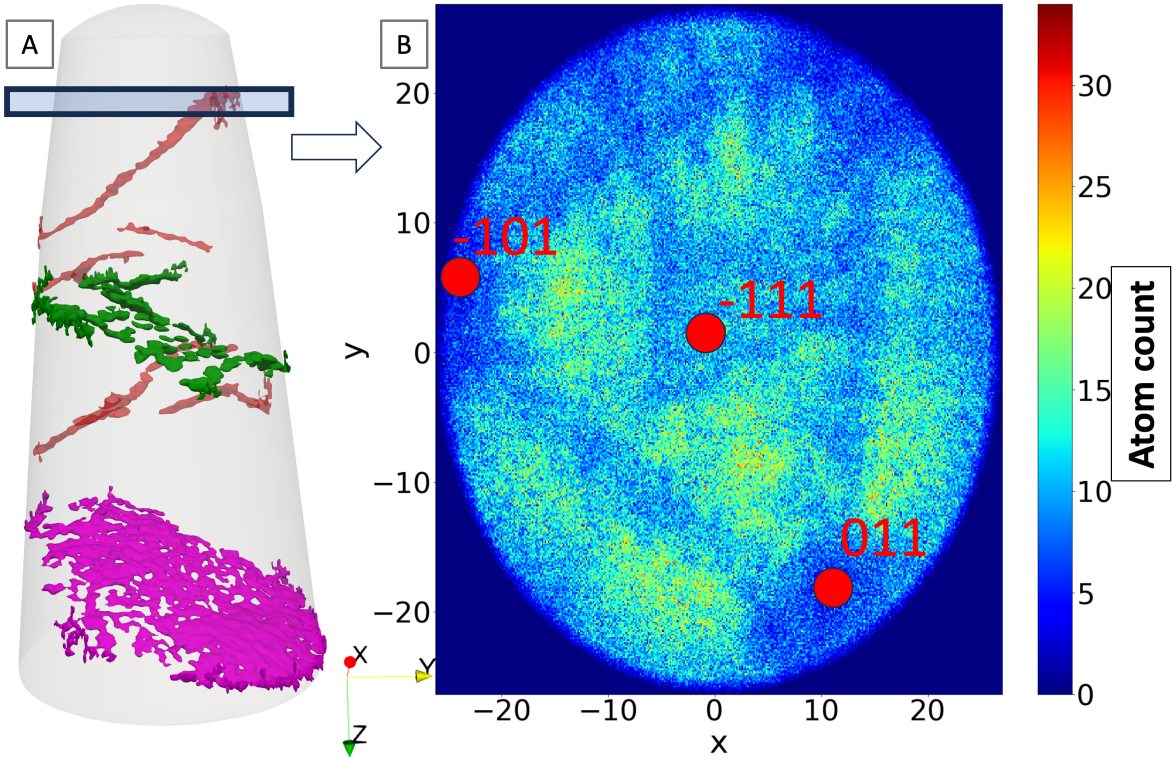



FIGURE 5.15: (A) Fe atoms within the blue rectangular slice are selected from the Fe-Mn alloy. (B) 2D histogram of the x-y coordinates of the selected atoms. Three low-density regions or poles are labeled according to the crystallographic planes they correspond to.

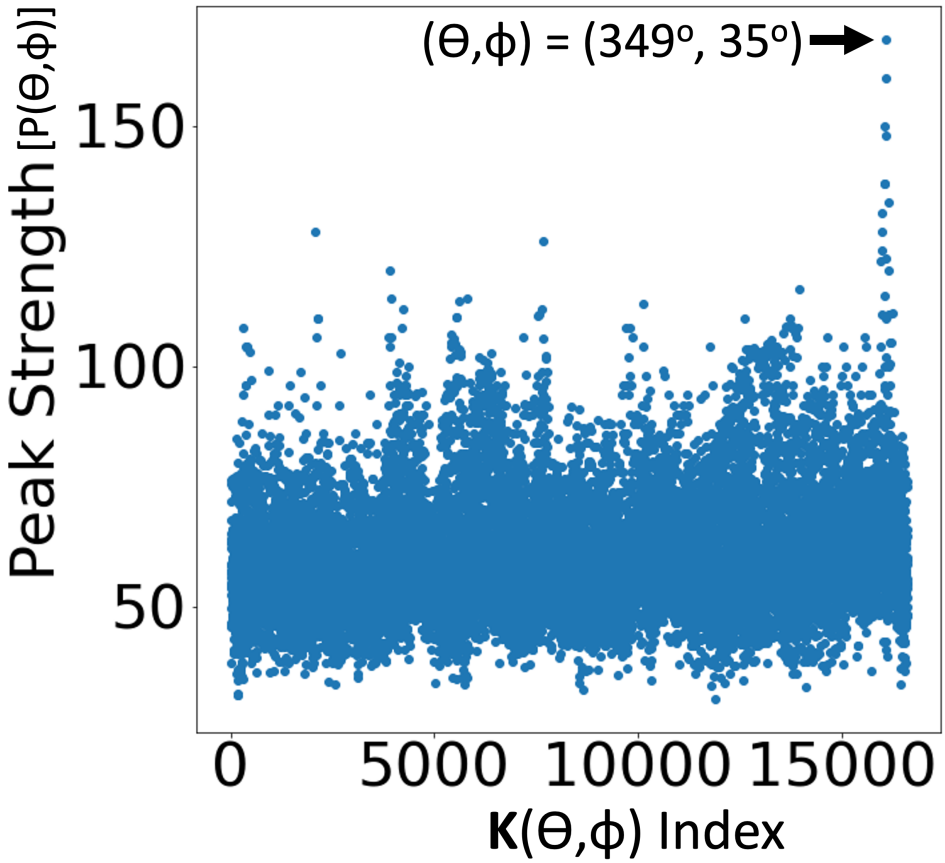



FIGURE 5.16: Peak strength or $P(\theta, \phi)$ values calculated after FFT analysis for vectors $\mathbf{K}(\theta, \phi)$, which are defined for ordered pairs of azimuthal $(\theta)$ and zenith angles $(\phi)$. The plot corresponds to a spherical ROI at the (-101) pole. The maximum value of $P(\theta, \phi)$ is pointed out and the corresponding values of $\theta$ and $\phi$ are also mentioned.

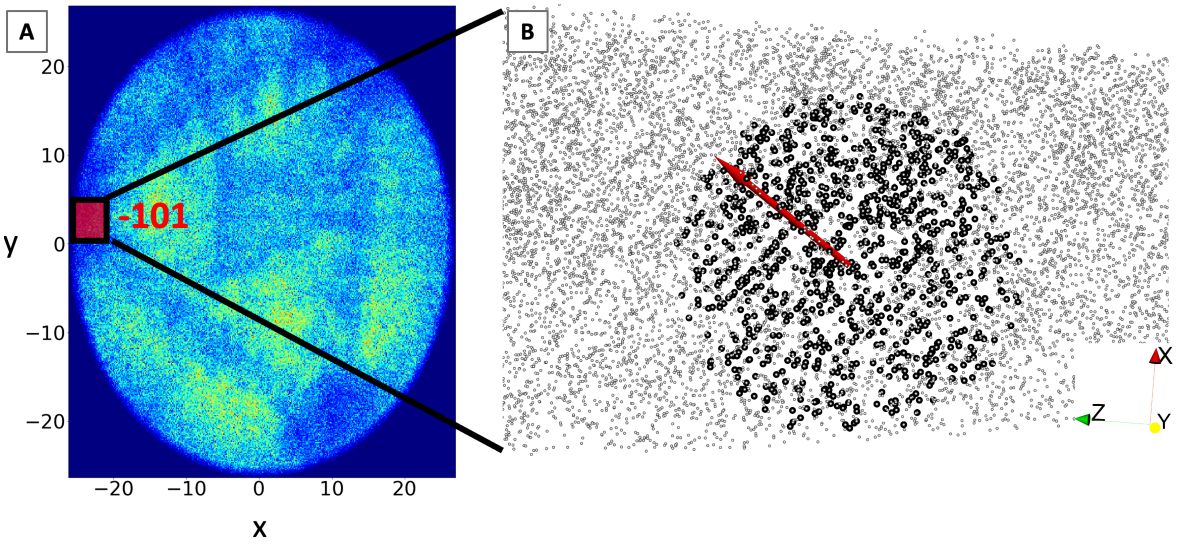



FIGURE 5.17: (A) Fe atoms around (-101) pole are selected. (B) After orientation analysis the vector $\mathbf{K}(\theta, \phi)$ (in red) with maximum value of $P(\theta, \phi)$ (refer Figure 5.16) is shown. $\mathbf{K}(\theta, \phi)$ is normal to the atomic planes.

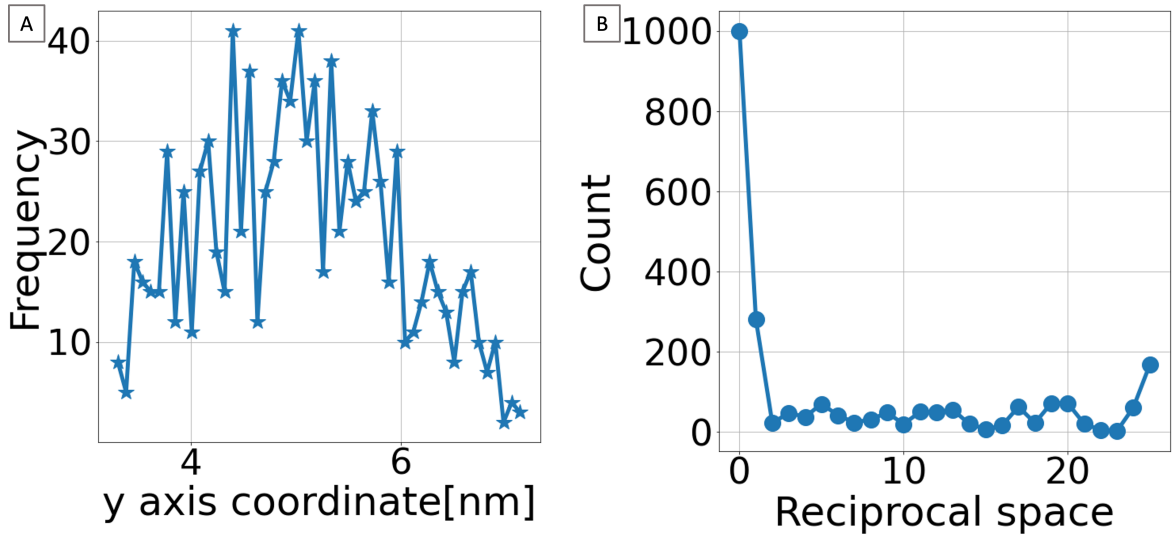



FIGURE 5.18: The vector $\mathbf{K}(\theta, \phi)$ that is normal to atomic planes at the (-101) pole is rotated to align with the y-axis, the histogram of the y-coordinates of the Fe atoms after rotation is shown in panel (A). The absolute values after applying FFT on the y-coordinate histogram are shown in panel (B).

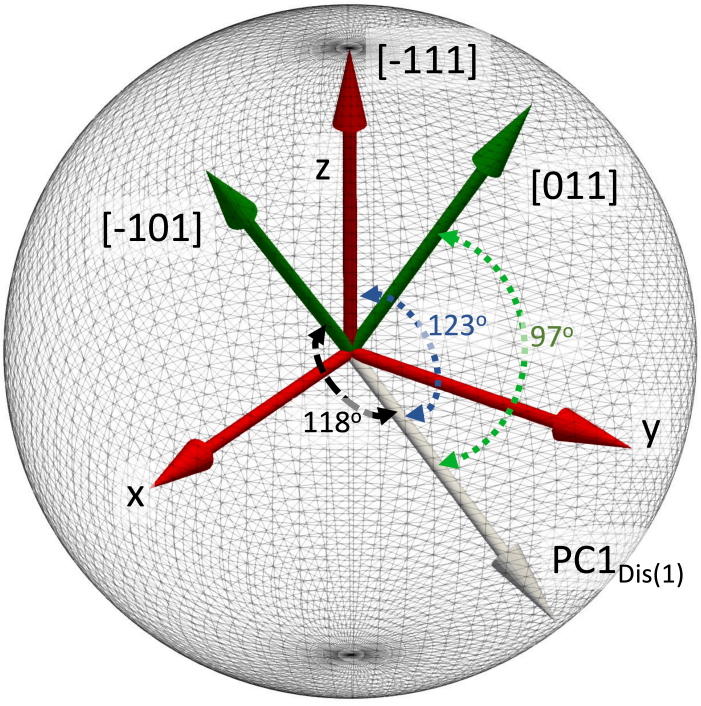



FIGURE 5.19: Sample coordinate system is shown in red color. The crystallographic directions [-101], [-111], [011] extracted after the orientation analysis are shown in green color. $PC1_{Dis(1)}$ corresponds to PC1 vector for the dislocation marked 1 in the Figure 5.14.

this step of isovalue selection is not done automatically and has to be set by the user. In the future, however, a method based on machine learning or statistical techniques can be developed to choose the most optimum isovalue for accurate microstructure description. In addition, in the present case, to find out the optimum normal to the crystal planes in orientation analysis, the algorithm tests a wide range of azimuthal and zenith angles, which is time-consuming if the orientation analysis is performed on whole APT reconstruction data. However, Bayesian optimization can be used to speed up the search for correct angles.

By enabling automatic segmentation and composition analysis, the workflow has the potential to provide insights into the role of dislocations in affecting the properties of a material. This approach not only decreases the human effort but also opens avenues for high throughput study analyzing property-structure relationships.

# Chapter 6

# Applying supervised learning to extract local ordering information from APT data

*The Al-Mg-Li alloy APT dataset was provided by Yue Li and Baptiste Gault.*

## 6.1 Introduction

In the previous chapter, crystallographic information is extracted from the APT data to calculate the orientation of dislocations in the crystal coordinate system. This was achieved by finding vectors in the reconstruction data coordinate system that are perpendicular to the atomic planes at the poles where spatial resolution in the depth direction is better than in the lateral direction. The correspondence between these vectors and the pole's Miller indices then helped to compute the orientation of any microstructural feature in the crystal coordinate system. Here, I have extended the idea of using crystallographic information from APT data to yet another step by using it to distinguish regions with ordered crystal structure using machine learning. The local ordering in a metallic alloy can significantly influence the mechanical and functional properties of the material [118]. Therefore, the development of tools for segmenting and identifying such regions, particularly in 3D reconstruction data, is crucial for better understanding and optimizing the material's performance [43], [118].

In APT reconstructed data, atomic planes can be resolved near the poles [119]. Many methods have been developed to analyze the spacing and periodicity of these atomic planes and correlate them to the atomic structure of the material being analyzed [33]–[35]. The methods that have been developed include using fast Fourier transforms (FFT) in a similar way I did in the previous chapter and radial distribution-based spatial distribution maps (SDMs) [35]. SDMs are traditionally used to analyze periodicity in a local region of APT data for identifying crystal structures or local ordering [118]. To make this identification automated, Li et al. used 2D SDMs to capture the atomic planes near a pole in the APT data and trained a convolution neural network (CNN) to learn patterns in the SDMs and distinguish $L1_2$-ordered domains from the FCC matrix [44]. Also, 1D SDMs were used by Li et al. in conjunction with 1D CNN to find chemical short-range order in the APT data [43]. 1D or 2D SDMs contain the distance information between the atoms; however, they don't capture the 3D environment around each atom and the directional (orientation) information. Recently, Rao et al. used 3D CNN on voxelized APT data, where each voxel contained the local composition information, to separate $L1_2$ structures from the FCC matrix [120]. This method captured the 3D information in the APT data. However, using 3D images of the APT data containing voxels/pixels is not an efficient way of including all the symmetries in a given crystal structure which means a lot of training

data has to be curated to train the machine learning model. This is because the 3D voxel images are not translationally and rotationally invariant. On top of this due to voxel resolution the atomic information crucial for classification might also be lost.

Smooth Overlap of Atomic Positions (SOAP) are rotationally and translationally invariant descriptors that encapsulate local structural and chemical environments around each atom in a given 3D atomic structure [121]. In recent years, they have been extensively used to fit interatomic potentials [121], [122]. Leitherer et al. used SOAP descriptors to represent various synthetic atomic structures and then trained a deep neural network (DNN) to learn the relationship between the structure label and SOAP representation [123]. The trained model was used to recognize different structures in experimental data pertaining to high-angle annular dark-field (HAADF) imaging, high-resolution transmission electron microscopy (HRTEM), and atomic electron tomography [123], [124]. The spatial resolution of these experimental techniques is far better compared to APT. Therefore, the limited spatial resolution poses a huge challenge when applying the method proposed by Leitherer et al. directly to the APT data.

In this chapter, I have used an autoencoder architecture to tune the hyperparameters of the SOAP descriptors so that they can capture the signature of local order within the noisy APT reconstructed data. The SOAP description of the synthetic data that emulates the spatial noise of the APT is used to train a DNN. The trained model is then used to classify $L1_2$ ordered domains in an Al-Mg-Li alloy. SOAP descriptors can represent any atomic structure within 3D APT data without relying on any voxelization and they implicitly contain the distance information conveyed by SDMs. Additionally, the rotational and translational invariance inherent in the SOAP description contributes to the development of more versatile machine learning models. I believe these models can be made capable of identifying atomic structures across different poles in given APT data or even across different APT datasets.

## 6.2 Methods and materials

### 6.2.1 Experimental data

An APT dataset corresponding to Al–5.18Mg–6.79Li (at.%) alloy has been used to showcase the developed method. The details related to APT measurements and sample preparation for this alloy are discussed in [118]. The alloy has been aged for 8 h at 150 °C. It consists of almost spherical precipitates of $\delta'$ phase in a random solid solution of Li and Mg in the Al FCC matrix. The $\delta'$ precipitates have approximately $Al_3$(Li, Mg) stoichiometry, $L1_2$ structure, and are fully coherent with the FCC matrix. In the $\delta'$ phase, Mg atoms are situated at the Li sublattice locations within the $L1_2$ structure. Hence, in the current investigation that involves differentiating between $L1_2$ and FCC structures in APT data, all the Mg atoms are replaced by Li atoms in the experimental data. In the context of experimental data from this point forward, both Mg and Li atoms are represented as Mg/Li.

### 6.2.2 Smooth Overlap of Atomic Positions (SOAP)

SOAP descriptors encapsulate the local structure and chemical environment around an atom in a given atomic structure [125]. The local neighborhood around the central atom for which the SOAP descriptors/vector is calculated is defined by selecting all the atoms within a cutoff radius $r_{cut}$. The local atomic density around the central atom is calculated by taking the sum of the Gaussian functions with standard deviation $\sigma_s$ centered at each neighboring atom and the central atom [122]. The Gaussian density is projected to orthonormal radial basis functions ($g_n$) and spherical harmonics

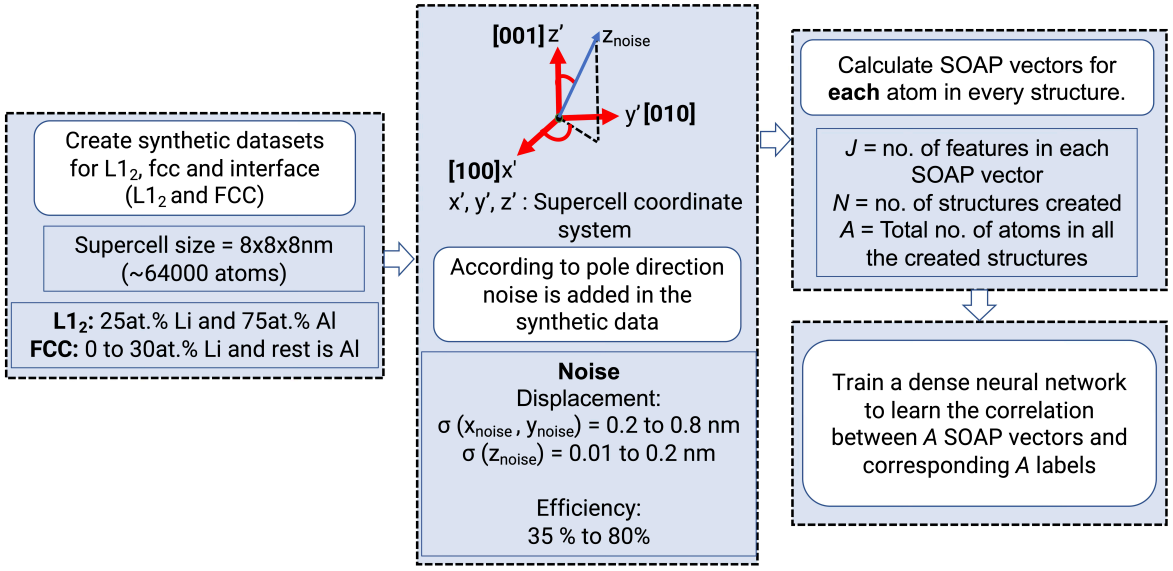



FIGURE 6.1: Workflow for curating synthetic APT data to train a dense neural network for automatic classification of $L1_2$ ordered domains.

($Y_{lm}$) to describe the local 3D environment around the central atom. The SOAP vector elements $p$ are given by

$$p_{nn'l}^{Z_1Z_2} = \pi \sqrt{\frac{8}{2l+1}} \sum_m c_{nlm}^{Z_1} {}^* c_{n'lm}^{Z_2} \tag{6.1}$$

where $Z_1$ and $Z_2$ are the atomic numbers of the chemical species in the system. $n$, $n'$, $l$ and $m$ are integers. $n$ and $n'$ correspond to the index of radial basis functions, with the maximum value being $n_{max}$. $l$ is the degree of the spherical harmonics whose maximum value is $l_{max}$ [122], [125]. The values of $n$, $l$, and $m$ ($= [-l, l]$) determine the level of detail or resolution in which the SOAP descriptors describe the atomic environment surrounding a specific central atom. The coefficient $c$ for chemical species $Z$ in equation 6.1 is

$$c_{n,l,m}^{Z} = \int d^3\mathbf{r}\; g_n(\|\mathbf{r}\|) Y_{lm}\left(\frac{\mathbf{r}}{\|\mathbf{r}\|}\right) \rho^{Z}(\mathbf{r}) \tag{6.2}$$

where $g_n$ is the radial basis function, $Y_{lm}$ are spherical harmonics, $\rho$ is the fitted Gaussian density, $\mathbf{r}$ is the vector from a given central atom to another atom within the cutoff radius [122], [125]. The SOAP vector captures the information about the local chemistry as well as local atomic orientation simultaneously around a given atom. In the present work, the SOAP vectors are calculated using DScribe, a python-based package used for computing atomic fingerprints/descriptors [125], [126]. To calculate SOAP descriptors around an atom hyperparameters $r_{cut}$, $n_{max}$, $l_{max}$, and $\sigma_s$ have to be provided by the user.

### 6.2.3 Learning local atomic environments in APT

Initially, synthetic data devoid of noise pertaining to the FCC and $L1_2$ structures is generated (see Figure 6.1). The perfect $L1_2$ ($Al_3Li$) structure has 25at.%Li and rest is Al. In each FCC supercell, the Li composition is arbitrarily selected within 0 to 30at.%, and the remaining is Al. The lattice constant for both structures is 4.07Å as in the given reconstruction dataset, the phases have a coherent interface. In the context of the given APT reconstruction data, $L1_2$ domains are embedded in the

FCC matrix. Thus, supercells encompassing the $L1_2$ and FCC interface are also generated. All the supercells are fabricated with non-periodic boundary conditions to ensure the correct classification of atoms situated at the edge of the reconstructed data. Also, to replicate the atomic environment inside the reconstructed data, larger supercells of size $8 \times 8 \times 8$ nm are chosen so that there are at least some synthetic atoms that are far away (>2nm) from the supercell boundary. Subsequent to this, noise is introduced to these datasets to emulate experimental APT data. Spatial noise is incorporated by displacing each atom by a distance, d, in a specified direction. This displacement follows a Gaussian distribution with a mean of 0 and a standard deviation of $\sigma$. Within APT data near the pole, the spatial resolution in the depth direction surpasses that in the lateral direction. Thus, for the synthetic data, the standard deviation in the depth direction, $\sigma_{z_{noise}}$, ranges between 0.01 nm and 0.2 nm, while in the lateral direction, $\sigma_{x_{noise},y_{noise}}$, it ranges between 0.2 nm and 0.8 nm, refer middle panel of Figure 6.1. To replicate the impact of APT detection efficiency, a portion of the atoms are randomly eliminated from the fabricated supercell. It's assumed that the detection efficiency ranges between 35 and 80% of the total atoms in a supercell, hence the remaining atoms are removed. Contrary to rotating the generated supercell to align crystal planes according to the pole direction, the noise axis $z_{noise}$ is rotated according to the pole direction, and pertinent noise is incorporated into the pristine supercell along this direction and the corresponding lateral directions (see middle panel of Figure 6.1). Given that SOAP descriptors exhibit rotational invariance, rotating the supercell or the noise axes yields identical results in atomic environments.

Pursuing the procedure above, N (here 16000) supercells are generated for FCC and $L1_2$ structures. Thereafter, SOAP vectors (of length J) for a defined set of hyperparameters, $r_{cut}$, $n_{max}$, $l_{max}$, and $\sigma_s$, are computed corresponding to each atom in every supercell. The label assigned to an atom is the same as that of its parent supercell, regardless of its chemical species. Consequently, a dataset comprising $A$ data points, where $A$ (here $\approx 10^6$) is the total number of atoms considering all the created structures, each with J features, is acquired.

This dataset serves as the input for a dense neural network encompassing L (here 6) hidden layers (see section 6.3.2). Each layer employs a rectified linear unit (ReLU) activation function, while the output layer utilizes a softmax layer to predict the probability of an atom residing in an FCC or $L1_2$ environment. The learning rate and batch size are 0.001 and 200, respectively.

### 6.2.4 Dimensionality reduction and hyperparameter tuning using autoencoders

Autoencoder is a neural network-based architecture that is used for dimensionality reduction [127]. Since SOAP descriptors can have more than a hundred components, it becomes difficult to interpret which features are most important for classification. A more general question would be what exactly the neural network model learns while classifying an environment as FCC or $L1_2$. This question comes under the realm of explainable AI. Due to the limited intuition associated with SOAP descriptors, the tuning of the SOAP hyperparameters for achieving high classification accuracy, which hinges on the effectiveness of SOAP in representing the local atomic environment, is also a cumbersome task. So, I have used autoencoders to reduce the dimensionality of SOAP space. Also, I demonstrate how autoencoders can help us understand what exactly neural networks are grasping from the SOAP space when categorizing atomic environments as FCC or $L1_2$.

The structure of an autoencoder is composed of two primary segments: the encoder and the decoder, as shown in Figure 6.2(A) [128]. Both the encoder and decoder sections have hidden layers of neurons. The data, represented as $\mathbf{X}$, is fed into the encoder. In the encoder, the number of neurons in each layer generally decreases until it reaches a predetermined number of dimensions (set by the user), known as the latent space. The decoder then uses the latent space as its input. For the decoder,

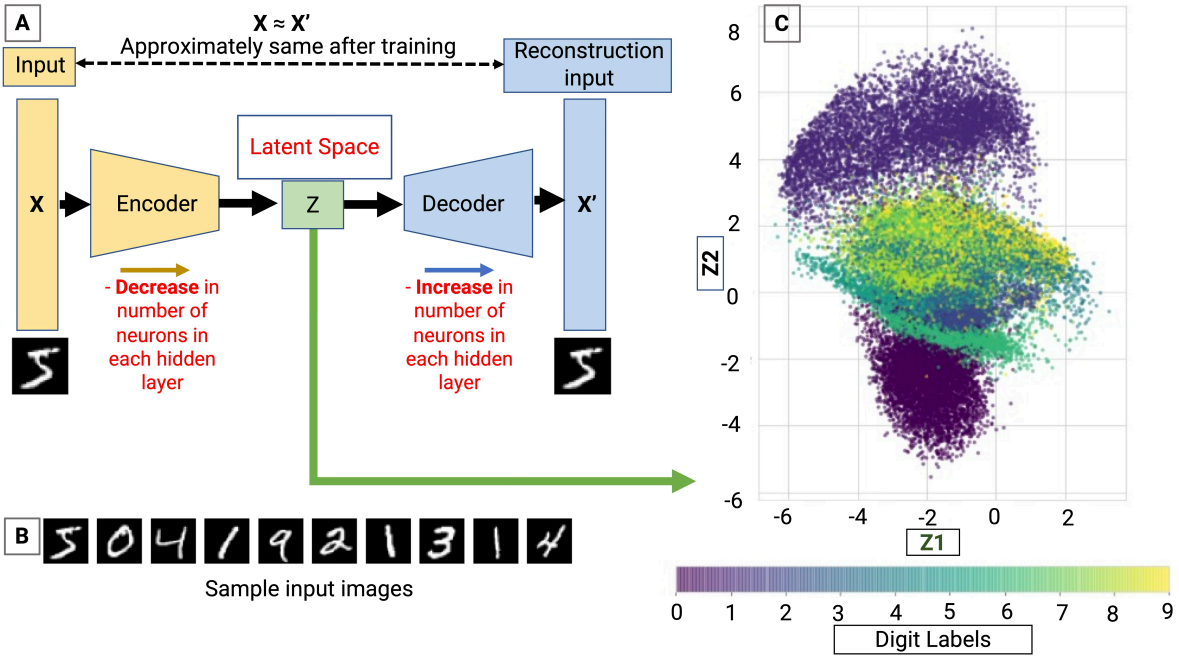



FIGURE 6.2: (A) Autoencoder architecture with 32,16 and 8 neurons in each consecutive layer in the encoder and 8, 16 and 32 neurons in each layer of the decoder. (B) Sample input images are fed to the encoder. (C) The two-dimensional latent space (Z) after training the autoencoder. Colors correspond to the label of each input image.

the number of neurons in the layers increases with each subsequent layer. The decoder's output, denoted as **X'**, retains the same dimensions as the encoder's input. The autoencoder aims to minimize the error between the decoder's output, **X'**, and the original data, **X**, essentially attempting to recreate the input data from the latent space. The encoder attempts to condense the information from the input data into the latent space, which usually has fewer dimensions compared to the input. Conversely, the decoder undertakes the more challenging task of reconstructing the input data from the latent space despite its reduced dimensions. Post-training, the latent space encompasses all the relevant information required to represent the given input dataset accurately.

To showcase the working of the autoencoder, it is applied to the well-known MNIST (Modified National Institute of Standards and Technology) dataset [129]. MNIST dataset is a pool of 60,000 images of handwritten digits from 0 to 9. So, there are ten classes in this dataset. Each image is in grayscale and has a size of $28 \times 28$ pixels, or there are 784 features describing each image. Some examples of the images comprising the MNIST dataset are given in Figure 6.2(B). This dataset is fed into an autoencoder. The details of the autoencoder are mentioned in Figure 6.2(A). The autoencoder tries to recreate the input image by minimizing the difference between its output image and the input. As shown in Figure 6.2(A) it tries to recreate the image of number 5. After training, the latent space, which is two-dimensional in this case, is basically the compressed form of the input images, as shown in Figure 6.2(C). Dimensions Z1 and Z2 in the panel (C) of the figure are latent space variables. So, the input image, which had 784 dimensions is now represented by two features, Z1 and Z2. The colors in Figure 6.2(C) depict the class of each image. It can be seen in the figure that images of the same handwritten digits are clustered together simply because although different people wrote them, they still have some similarities. However, there is overlap between the clusters of different classes, making it difficult to distinguish them.

To get a more structured or organized latent space, the autoencoder architecture is trained with a

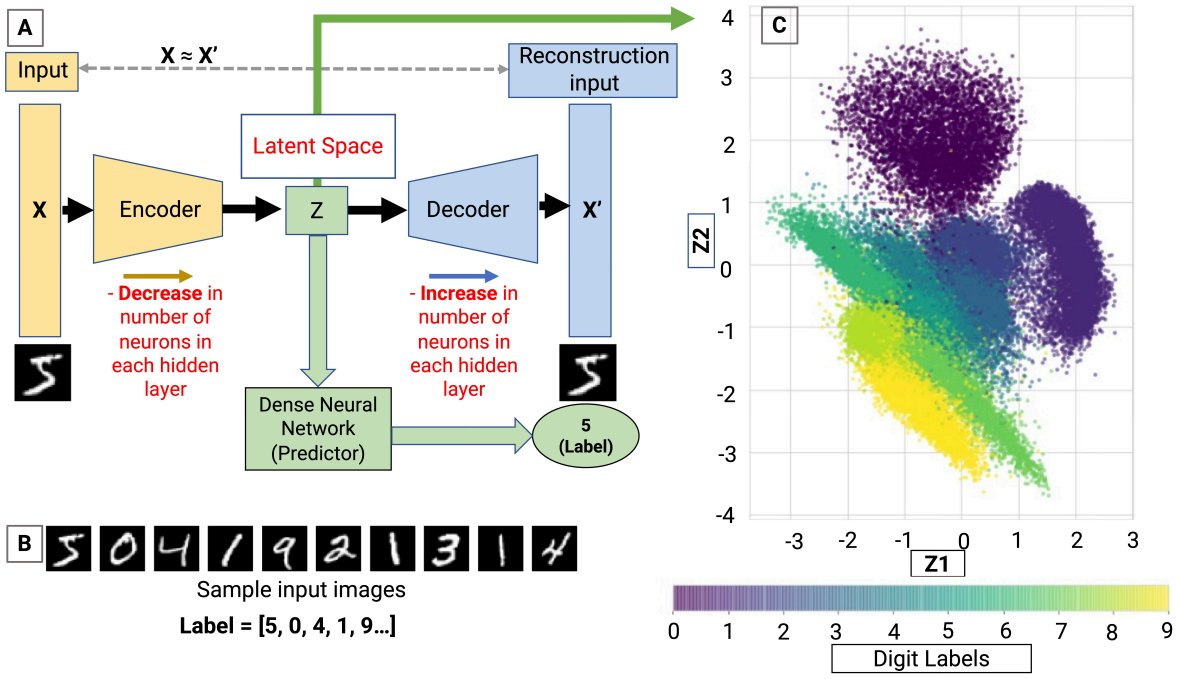



FIGURE 6.3: (A) Autoencoder architecture with a predictor (dense neural network). (B) Sample input images are fed into the encoder, and the corresponding labels are used to train the predictor. (C) The two-dimensional latent space (Z) after training the autoencoder-predictor architecture. Colors correspond to the label of each input image.

predictor, as shown in Figure 6.3(B) [128]. A predictor is a dense neural network for which the input data is the latent space, and the target variable is the labels of the handwritten digits. The predictor tries to find the correlation between the latent space representation of the input data and its label, which in this case is a number. Both the autoencoder and the predictor are trained simultaneously, which means that for the whole architecture, a common loss function is defined. After training this architecture on the MNIST dataset and the labels for each image, the latent space is shown in Figure 6.3(B). The latent space is more organized according to labels, with each cluster now clearly visible.

## 6.3 Results and discussion

From the Al-Mg-Li alloy at hand, the atoms near the (011) pole are extracted and reconstructed. The Mg/Li atoms in the reconstructed data are shown in Figure 6.4(A). The high-density regions/clusters of Mg/Li atoms, visible in the image have $L1_2$ ordered structure. The matrix phase has FCC structure in which solute atoms (Mg/Li) are randomly substituted. The schematic FCC and $L1_2$ supercells are depicted in Figure 6.4(C and D). The supercells are rotated such that the [011] direction of the supercell is parallel to the z-axis of the reconstructed data. Figure 6.4(B) shows the magnified projection of Al atoms in the z-y plane. As mentioned previously, the resolution in APT reconstructed data is better in depth (z-direction) compared to the lateral direction(x-y). Therefore, atomic planes perpendicular to the z-axis can be seen in the magnified image. In the upper left corner, marked by a green box, the atomic planes of Al appear to have more distance between them compared to the planes in the lower right corner (red box). This is because the upper left region contains the $L1_2$ phase and the lower region comprises the FCC phase.

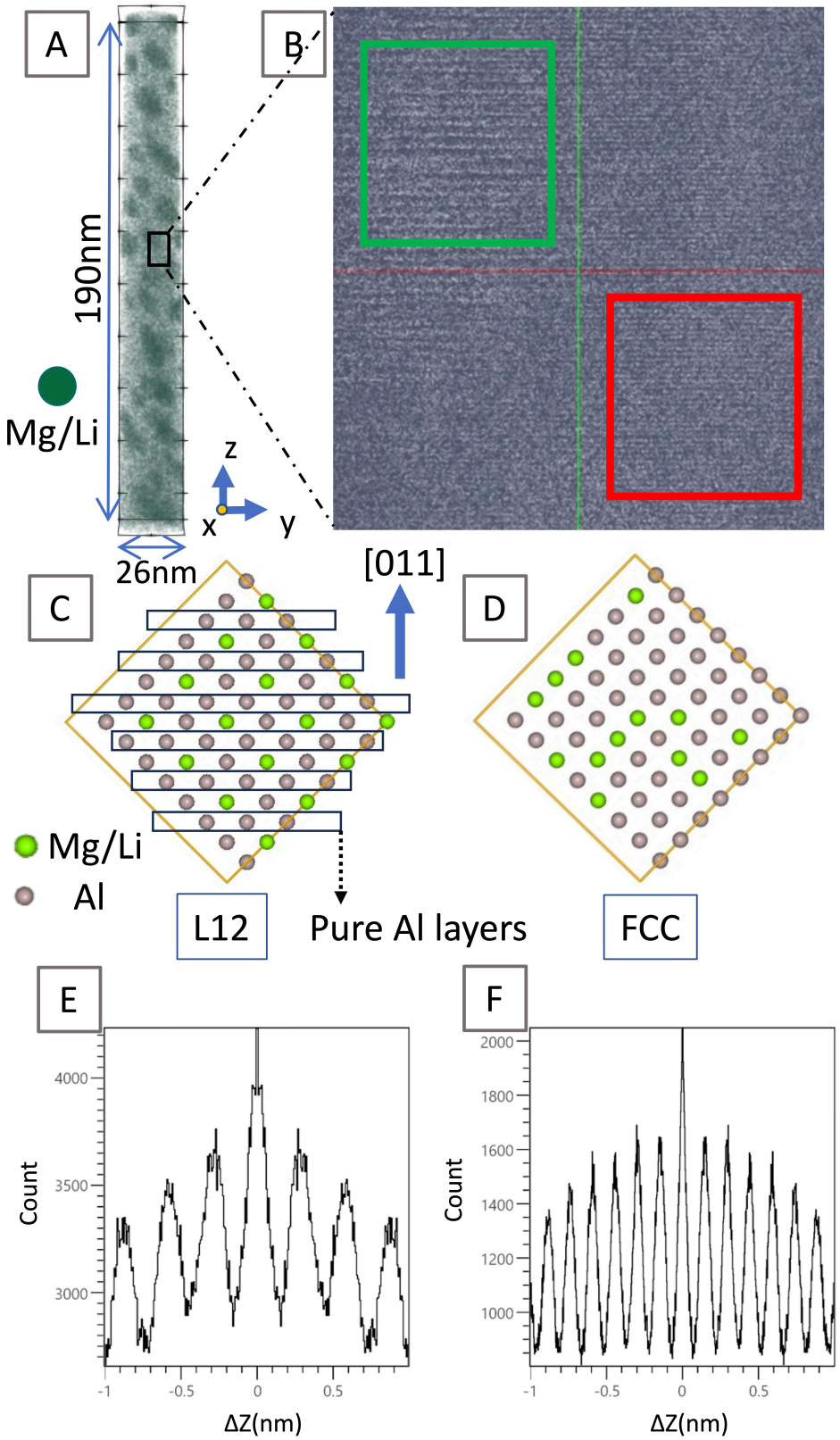



FIGURE 6.4: (A) Mg/Li atoms in the reconstruction data corresponding to the (011) pole in Al-Mg-Li alloy. (B) Magnified view of the regions highlighted by a black box in panel (A). The green box refers to the $L1_2$ region, and the red box corresponds to the FCC matrix. (C and D) synthetic $L1_2$ and FCC supercells rotated such that the [011] direction aligns with the z-axis of the reconstruction data. In (C) the black boxes highlight the pure Al layers. (E and F) z-SDMs (Al-Al) for $L1_2$ and FCC regions in the reconstruction data.

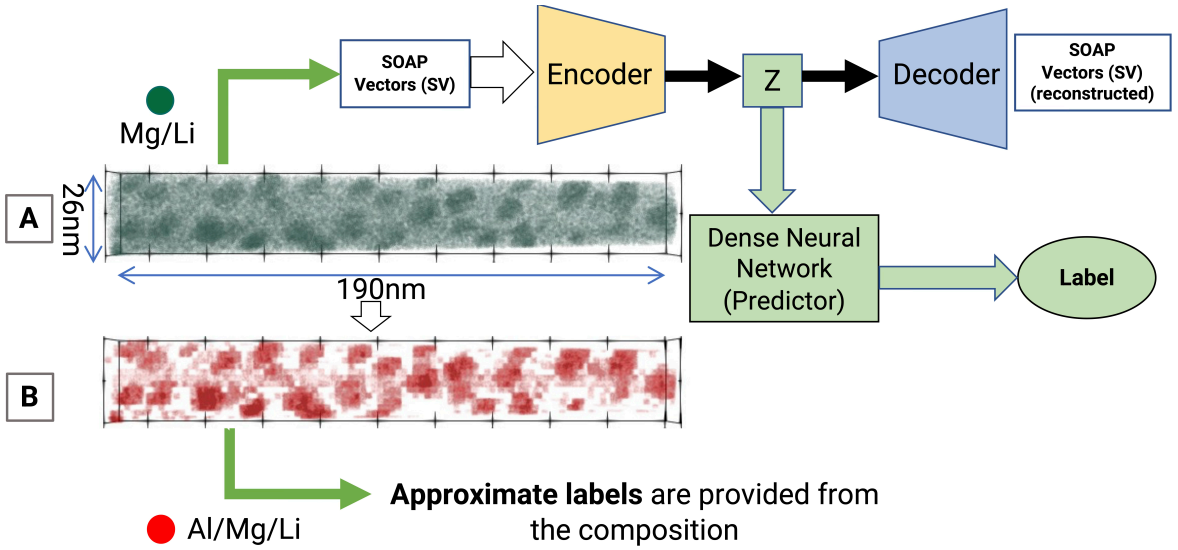



FIGURE 6.5: (A) SOAP vectors of all the atoms (Al and Mg/Li) in the reconstruction data are utilized for training an autoencoder and simultaneously the labels of the atoms are used to train the predictor. (B) The approximate labels of the atoms for training the predictor are provided by labeling atoms in the voxels with a composition higher than 10at.% Mg/Li as $L1_2$, and rest are labeled as FCC.

The $L1_2$ phase has an ordered structure, so moving in the [011] direction (or Z axis ), we will find that each alternate layer has only Al atoms. Black boxes also highlight these pure Al layers in the schematic $L1_2$ supercell in Figure 6.4(C). On the contrary, in the FCC phase, each layer can have solute atoms randomly placed. Therefore, in the magnified image, Al layers in $L1_2$ are more spaced out compared to those in the FCC phase. This can also be proven by plotting z-direction SDMs for the Al atoms in the respective $L1_2$ and FCC regions. As shown in Figure 6.4(E and F), the Al-Al peaks are more apart in the $L1_2$ phase than in z-SDMs for the FCC phase.

Here, the aim is to capture the local structure and chemical environment around each atom in the experimental data using SOAP descriptors and then classify each atomic environment represented by SOAP as either FCC or $L1_2$ using a trained dense neural network model.

### 6.3.1 Autoencoder application to the experimental data

SOAP descriptors have not been used before to describe atomic environments in APT data which is very noisy compared to the pristine crystalline structure. Therefore, the SOAP hyperparameters ($r_{cut}$, $n_{max}$, $l_{max}$, and $\sigma_s$) that are optimum to describe such a noisy data are not known. One can imagine the concept of SOAP as a camera, the hyperparameters as the length between the lens of the camera and the optical sensor. For a sharp image the light must converge at the sensor which can only happen if the distance between the lens and sensor is optimum. Similarly, only optimum SOAP hyperparameters can describe the atomic environments accurately without noise. However, as mentioned before SOAP vectors are difficult to interpret because of their higher dimensionality. So, the autoencoder with predictor architecture is used to reduce the dimensionality of the SOAP space to find the optimum hyperparameters for $L1_2$ and FCC classification (see Figure 6.5).

The SOAP vectors of each atom, irrespective of their chemical nature, in the reconstructed data are extracted for given values of $r_{cut}$, $n_{max}$, $l_{max}$, and $\sigma_s$. The dimensionality of SOAP space depends on the hyperparameters used to calculate it. In this data set the $L1_2$ domains are richer in solute atoms compared to the matrix phase. The reconstruction data is voxelized and the atoms residing in voxels with composition higher than 10at.% Mg/Li are labelled as $L1_2$ phase, see Figure 6.5(B). As shown

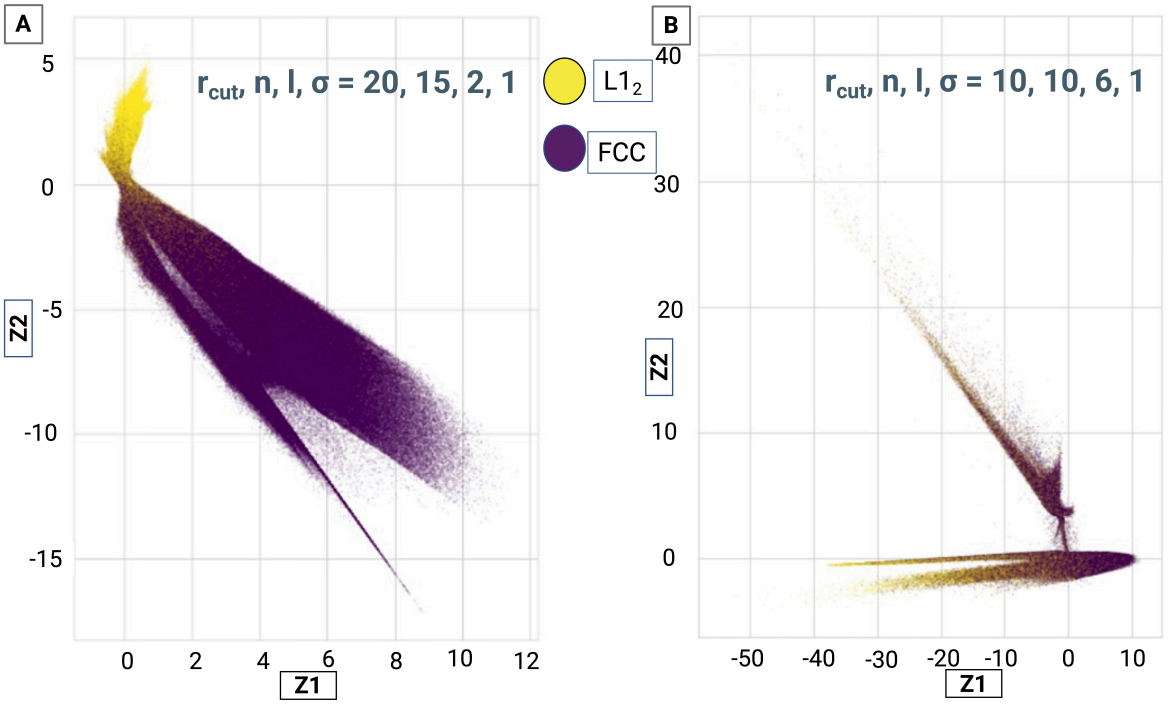



FIGURE 6.6: Latent space for two different sets of SOAP hyperparameters after training the autoencoder-predictor architecture given in Figure 6.5(A) on the reconstruction data. SOAP hyperparameters [$r_{cut}$, $n_{max}$, $l_{max}$, $\sigma_s$] are H1 = [20, 15, 2, 1] in (A) and H2 = [10, 10, 6, 1] in (B).

in the image, this classification is approximate and not very accurate as some voxels even outside $L1_2$ clusters are labeled as $L1_2$. This approximate labeling or noise will prevent the architecture from getting overfitted. The SOAP vectors extracted from each atom are fed into the encoder part of the autoencoder. Also, for each SOAP vector, the labels (FCC/ $L1_2$) calculated above are provided to the dense neural network/predictor as the target variable (see Figure 6.5(A)). During training, the autoencoder tries to reproduce the SOAP vector for each atom and the predictor tries to learn the correspondence between the latent space representation of the SOAP vector and its label.

The autoencoder predictor architecture is trained on sets of SOAP vectors calculated with the following hyperparameters: H1 = [20, 15,2,1] and H2 = [10,10,6,1]. The values in H1 and H2 correspond to $r_{cut}$, $n_{max}$, $l_{max}$, and $\sigma_s$ respectively. After training the SOAP latent space is shown in Figure 6.6(A and B) for the two sets of hyperparameters. Each point in the latent space corresponds to an atom in the reconstruction data. The colors in the figures depict the atomic environment (FCC/$L1_2$) around each atom. Clearly, for the H1 hyperparameters in the latent space, the atoms of each class are clustered together. On the contrary, for H2 hyperparameters, two clusters are observed but they do not correspond to the two classes/labels, see Figure 6.6(B). So, the first set of hyperparameters is accurately able to capture the structure and chemistry around each atom to ensure accurate classification and separation of atoms in latent space.

Furthermore, the SOAP vectors with H1 hyperparameters are calculated, and the SOAP vectors of the atoms that lie inside the reconstruction dataset are selected, as shown in Figure 6.7(A). The atoms that lie at the edge of the dataset might have slightly different SOAP vectors due to non-periodic boundary condition. Selecting the atoms inside the dataset removes the SOAP vectors with edge effect. The latent space for the autoencoder-predictor network trained on SOAP vectors for atoms present inside the dataset is shown in Figure 6.7(B). As seen in the figure, the latent space is partitioned into two clusters corresponding to the two classes. The overlap between the clusters can also be attributed

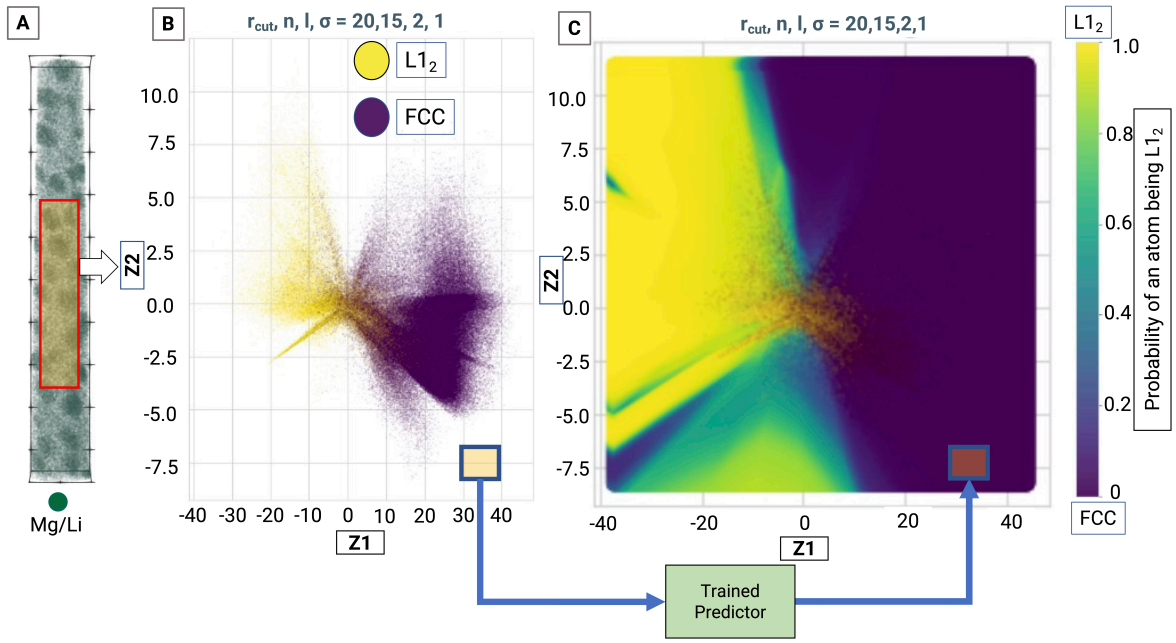



FIGURE 6.7: (A) SOAP vectors (calculated for H1 hyperparameters) of the atoms inside the red box are selected. (B) Latent space after training an autoencoder-predictor architecture on the selected SOAP vectors. (C) The trained predictor predicts the probability of being in $L1_2$ environment for the whole latent space.

to the fact that the initial classification of atoms as FCC and $L1_2$, using voxel-based partitioning, was approximate.

The partitioning of the latent space into FCC and $L1_2$ can be made more continuous. For example, in Figure 6.7(B) there are no datapoints present in the latent space region circumscribed by the blue rectangular box. The Z1 and Z2 variables from this region are fed into the trained predictor, which predicts the labels of the initially empty rectangular region, see Figure 6.7(C). Following this procedure, the labels for the whole latent space are predicted to obtain a continuous partitioning between the two classes. Clearly, the continuous latent space is partitioned between FCC and $L1_2$ regions. However, there are regions that had a smaller number of input datapoints that do not explicitly belong to a single class (probability of being $L1_2$ is between 0.2 and 0.8). The latent space in Figures 6.6(A) and 6.7(C) shows that the H1 hyperparameters can be used to encapsulate the local structure for classification of $L1_2$ and FCC regions in APT data. The continuous latent space partitioning in Figure 6.7(C) demonstrates what features the neural network/predictor learns for classifying an atomic environment as FCC or $L1_2$ even for regions without input data.

### 6.3.2 Neural network application on the experimental data

A synthetic APT dataset containing FCC and $L1_2$ supercells is created, following the procedure mentioned in section 6.2.3. The SOAP vectors corresponding to H1 hyperparameters are extracted for each atom in every synthetic supercell. Since H1 is a set of optimized hyperparameters the corresponding SOAP vectors can capture the atomic environments accurately. The label of each atom/SOAP vector is same as its parent supercell. The SOAP vectors and the corresponding labels are divided into train (75%) and test datasets (25%). A dense neural network with six hidden layers (other hyperparameters related to DNN are mentioned in the section 6.2.3) is trained until 200 epochs (see Figure 6.8(A)). The training and testing loss after each epoch is shown in Figure 6.8(B). As elucidated in the figure, in the first few epochs, the training and testing loss drop sharply, depicting that the

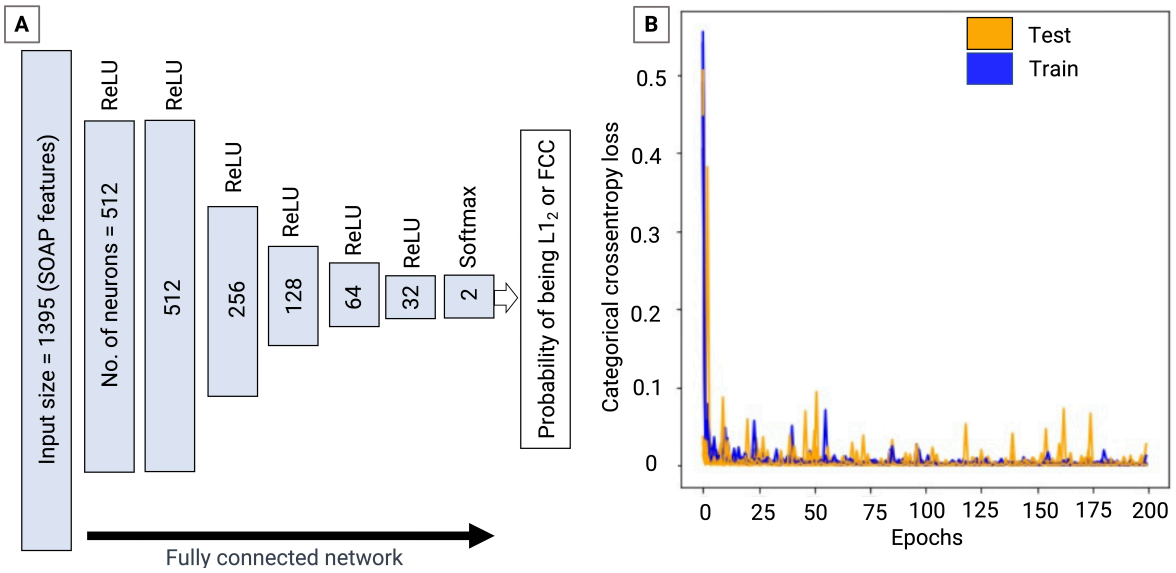



FIGURE 6.8: (A) Dense neural network architecture trained to get the correlation between the synthetic data SOAP vectors and the corresponding labels (atomic environment). After each hidden layer, there is a dropout layer. The two neurons in the output layer predict the probability of an atom being in the $L1_2$ and FCC environments. (B) Test and Train loss for each epoch while training the DNN.

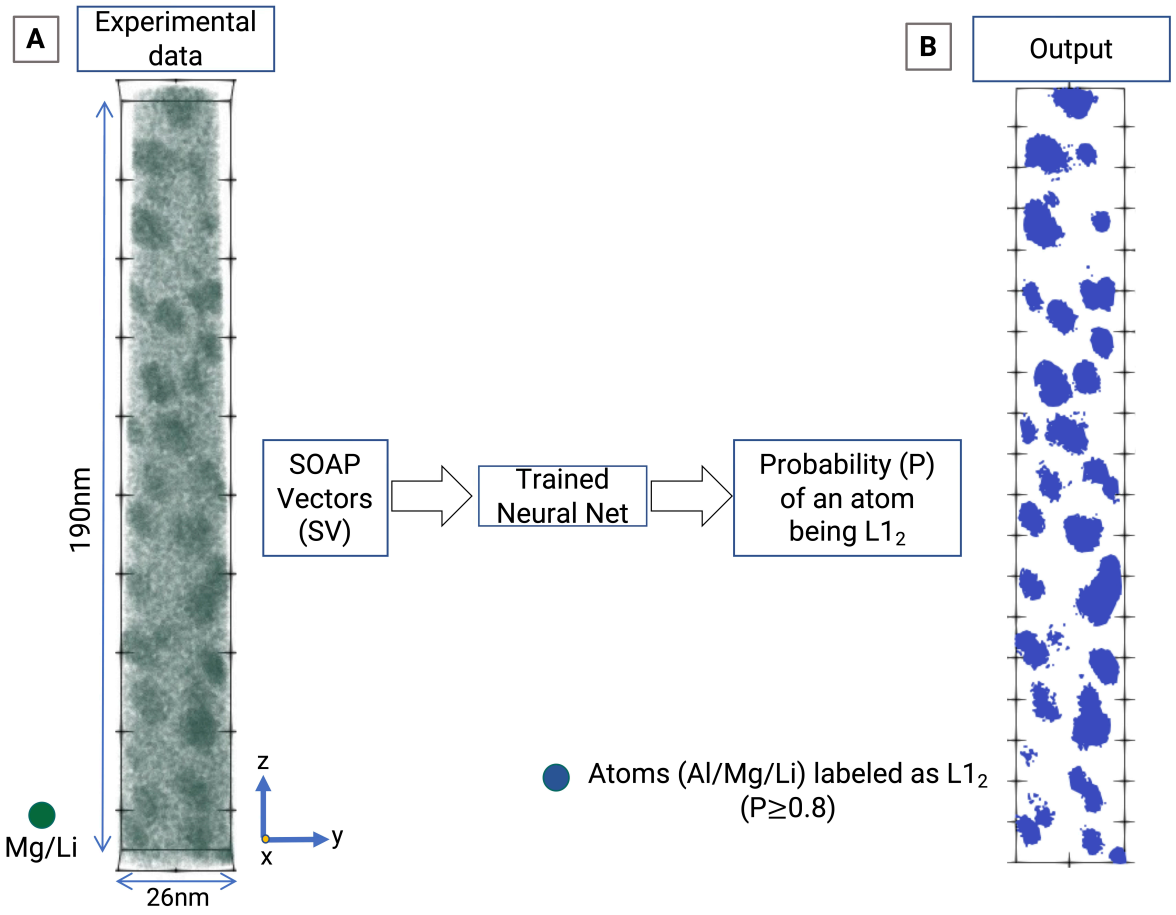



FIGURE 6.9: SOAP vectors corresponding to the reconstruction data are fed into the trained neural network to predict the probability of the atoms being in the $L1_2$ atomic environment. Atoms with the probability of being $L1_2 \geq 0.8$ are shown in (B). The atom clusters in (B) visually match the clusters of Mg/Li with $L1_2$ structure in the reconstruction/experimental data in (A).

model is learning the correlation between SOAP vectors and the labels. Further, the training loss is not lower than the testing loss, depicting that the model is not overtrained. The fluctuations in the test and training loss are attributed to the fact that the samples in a batch for training are selected randomly and also because there is a drop-out layer after each hidden layer to avoid overtraining. After training the model, we get an accuracy of 99.8% on the test dataset. This trained model is then used to find the label or the atomic environment around each atom in the reconstructed dataset as shown in Figure 6.9. So, first, the SOAP vectors of each atom in reconstructed data are calculated using H1 hyperparameters. The calculated SOAP vectors are fed into the trained DNN to predict the labels (FCC/$L1_2$) of each atom. The atoms classified as $L1_2$ (probability $\geq 0.8$) are shown in Figure 6.9(B). These clusters of atoms (Al, Mg/Li) can be compared to the Mg/Li atom clusters that correspond to the $L1_2$ domains in the reconstructed data shown in Figure 6.9(A). The model successfully identifies the $L1_2$ phase domains based on the 3D local structure and chemistry around the atoms in the reconstructed data.

## 6.4 Conclusion

In summary, the chapter illustrates the enhancement of machine learning models for identifying atomic structures within 3D APT data by employing SOAP descriptors. The adjusted autoencoder architecture was critical in tuning the SOAP descriptors to represent the noisy APT reconstruction data aptly. The training of a DNN using synthetic data emulating the spatial noise of APT data helped in the accurate classification of $L1_2$ ordered domains within the Al-Mg-Li alloy. Unlike conventional methods that hinge on voxelization or SDMs, which may lead to loss of vital atomic information and require extensive training data due to lack of translational and rotational invariance, the utilization of SOAP descriptors eliminates such dependencies, offering a more robust and efficient avenue for machine learning applications in APT.

The rotational and translational invariance inherent in SOAP descriptors not only alleviates the need for cumbersome data curation but also paves the way for developing more versatile machine learning models. These models bear the potential of identifying atomic structures across varying poles within a given APT data or even across different APT datasets. However, this requires better strategies for curating training datasets that can include APT artifacts than those presented in this chapter. One way of training data curation is to perform physics-based simulations that can capture APT artifacts and corresponding noise in the data. Additionally, within this chapter for the Al-Mg-Li dataset, the hyperparameters of SOAP were fine-tuned utilizing autoencoders; nonetheless, it's possible that the identical hyperparameters may not be effective for other material datasets.

This chapter shows a promising way to use SOAP descriptors-based machine learning models to extract and analyse crystallographic information from APT data, therefore paving the way for better understanding and tailoring material properties.

# Chapter 7

# Summary and Outlook

This thesis focuses on developing machine learning (ML)-based workflows to automate the analysis of atom probe tomography (APT) datasets and extract as much information as possible from these datasets. APT involves sequential field evaporation of atoms from the surface of a needle-shaped specimen to retrieve the location of the hit on a 2D position-sensitive detector and the time of flight information of each detected ion. These are then used to provide a 3D distribution of atoms within the material with their chemical identities, enabling the visualization and analysis of microstructural features. Most of the analysis of the various microstructural features in APT data is still performed manually; therefore, automatic ML-based workflows are needed to be developed to ensure reliable and fast quantitative analysis.

In Chapter 3, an unsupervised ML-based method called clustering in the composition space-based approach, utilizing Gaussian mixture models and DBSCAN clustering algorithms, was developed to automatically segment out microstructural features using their chemical fingerprint from reconstructed APT data. The developed workflow separates each microstructural feature belonging to a phase or chemical domain as a separate entity. So effectively, there are two levels of segmentation first is segmentation based on composition which can differentiate between various phases or chemical domains, and the second is segmentation based on spatial distribution of the phase, for example, different precipitates or chemical subdomains of the same phase. This enables each entity's detailed compositional and geometric (for example, orientation, shape, or thickness) analysis. Notably, the workflow's versatility and efficiency were exemplified in its application to the Fe-doped Sm-Co alloy APT dataset. This demonstrated not only its rapid execution capability on large datasets ($\approx$ 500 million ions) but also its proficiency in segmenting complex microstructures. The workflow correctly identified the phases, namely matrix (2:17 phase), Z phase, and 1:5 phase, that are known to exist in the given Sm-Co alloy. The segmented microstructure for the alloy extracted after the analysis was visually identical to the microstructure delineated by the isosurfaces. However, challenges were encountered with the Ni-based superalloy, where the developed method could not accurately predict the number of phases in the alloy due to its multidimensional composition space. This led to the integration of a random forests algorithm-based clustering technique, enhancing the workflow's capability to handle APT datasets with many chemical species ($\geq$ 10). This adaptation added more flexibility to the developed method, making it suitable for high-throughput APT data analysis. Further, the composition space-based approach accurately estimated the volume of a secondary phase precipitate in a synthetic APT dataset compared to the traditionally used isosurfaces. Unlike isosurfaces, which rely strongly on user-defined isovalues, this method ensures more reliable and automatic segmentation of microstructural features in APT datasets. The limitations of the methodology, particularly in scenarios with subtle compositional contrasts between chemical domains, highlight areas for future refinement. Such challenges indicate the need for continuous evolution of the workflow to address diverse and complex APT datasets more effectively.

In Chapter 4, a novel APT data analysis pipeline was developed which uniquely addresses the challenge of quantifying the geometry and composition of morphologically complex microstructures, like agglomerates with plate-like substructures. This pipeline relies on the segmentation algorithm of Chapter 3 to effectively segment the microstructure, extracting key features such as volume fraction and average phase/chemical domain composition. The developed local PCA approach and U-Net based ML model are employed to reduce morphologically complex microstructural features into plate-like subdomains for further analysis. When applied to six APT datasets corresponding to the Fe-doped Sm-Co alloy, the pipeline identified three distinct phases and their compositional variations, crucial in understanding coercivity differences in the alloy. Notably, the pipeline discerned compositional differences in the 1:5 phase and Cu distribution in the Z-phase between regions of different coercivity. These insights, derived from the sophisticated analysis of the microstructure's composition and geometry, underscore the pipeline's capability to analyze complex microstructures in APT datasets, particularly those involving grain boundaries, phase boundaries, or planar precipitates. The pipeline thus stands as a powerful tool for consistent and detailed APT dataset analysis, offering significant advancements in the field.

Chapter 5 outlines a novel workflow for automatically extracting and analyzing line-like features, particularly dislocations in APT reconstruction data, enhancing existing methodologies. The workflow comprises several key steps: extracting and filtering isosurface meshes pertaining to dislocations using PCA, computing mesh skeletons through mean curvature flow, segmenting to isolate dislocation segments, fitting B-splines for precise representation, and placing regions of interest (ROIs) on the splines for detailed composition analysis. Additionally, it integrates crystallographic information from APT data for orientation analysis of dislocations in the crystal coordinate system. This kind of analysis is traditionally performed manually; for example, Dubosq et al. [130] had to manually find out the crystallographic planes/directions and dislocation line direction in the APT data to find the orientation of the dislocation in the crystal coordinate system. However, the developed workflow completely automates these various cumbersome steps. The workflow was applied to segment and analyze individual dislocations from a dislocation network in a Ni-based superalloy. Similarly, several dislocations were automatically extracted and studied for composition and orientation in a Fe-Mn alloy. Both case studies showed the robustness of the developed pipeline while dealing with complex linear features. While robust and largely automated, the workflow's effectiveness hinges on the user-defined selection of isosurface meshes, a step that could be optimized in the future using machine learning or statistical techniques for more accurate microstructure characterization. Furthermore, the current orientation analysis, though comprehensive, is time-intensive and could be expedited using Bayesian optimization [131]. This advanced workflow promises significant advancements in understanding the impact of dislocations on material properties, reducing manual effort and facilitating high-throughput studies that link structural and property relationships in materials.

Chapter 6 highlights the advancement of machine learning models for identifying atomic structures in 3D APT data, notably through the use of smooth overlap of atomic positions (SOAP) descriptors. The implementation of an autoencoder architecture was pivotal in effectively adapting SOAP descriptors to the spatially noisy APT data. This approach, especially in the context of classifying $L1_2$ ordered domains in Al-Mg-Li alloys, offers a general description of the 3D atomic environment over other methods reliant on voxelization or spatial distribution maps (SDMs), which can lose critical atomic information and demand extensive training data due to their lack of translational and rotational invariance. SOAP descriptors bring the advantage of inherent rotational and translational invariance, thereby reducing the need for extensive data curation and enabling the development of more versatile ML models. These models can be made capable of identifying atomic structures across different

poles within APT data or across various datasets. However, the chapter also identifies the need for improved strategies in curating training datasets to account for APT artifacts. Suggested methods include physics-based simulations to better represent APT noise and artifacts. Additionally, while the SOAP hyperparameters were effectively fine-tuned for the Al-Mg-Li dataset using autoencoders in the chapter, the hyperparameters for this example might not universally apply to other material datasets. Overall, this chapter presents a significant step forward in utilizing SOAP descriptors-based ML models for extracting and analyzing crystallographic information from APT data. This advancement opens up new possibilities for a deeper understanding and optimization of material properties.

## 7.1 Outlook

**Utilization of SOAP for atomic structure extraction from APT Data** : In this study, a method has been developed for generating synthetic data to train an ML model based on the Smooth Overlap of Atomic Positions (SOAP) descriptors aimed at extracting atomic structures from APT data. Currently, this method involves the rotation of each generated supercell to align with a specific pole or crystallographic direction identified in experimental data. However, a limitation is observed: the resulting model struggles to accurately identify atomic structures in reconstructed data at different pole directions. A potential solution has been identified, given the rotationally and translationally invariant nature of SOAP descriptors. A comprehensive training dataset encompassing a wide range of supercell rotations corresponding to different pole directions could be constructed. This approach again raises the necessity of optimizing SOAP hyperparameters, potentially using autoencoders. An alternative, more promising method for creating accurate synthetic data is the simulation of APT experiments using tools like TAPSim [132]. This simulation-based approach is likely to yield higher quality synthetic data than the current method, which involves introducing random noise to supercells.

**Enhancing the skeletonization workflow's robustness**: The effectiveness of the proposed skeletonization-based analysis in detailing linear features in APT data relies heavily on the isosurface meshes which represent the microstructure. The isosurface description of the microstructure in turn depends on the user-defined isocomposition values. To automate this process, composition space plots could be utilized. In these plots, the chemical domains/phases are segmented using the composition space-based method, assisting in identifying optimal isocomposition values through statistical or ML models. Additionally, while thorough, the current orientation analysis within the skeletonization workflow is time-consuming. Speeding up this process could be achieved by implementing Bayesian optimization [131].

**Using U-Net-based approach to analyze grain boundary segregation**: In this work, although not shown explicitly, the U-Net-based approach can also separate grain boundary networks, where solute atoms have segregated, into planar features. The inplane composition fluctuations within these planar areas can be quantified and examined for phenomena like spinodal decomposition, similar to the analysis performed by Zhou et al. in [46] for columnar grain boundaries.

**Automated thermodynamic analysis through clustering in composition space-based approach in APT Data**: Clustering within the composition space reveals the composition distribution for each phase or chemical domain in APT data. This allows for an automatic comparison of the average composition of each phase with the equilibrium compositions calculated using CALPHAD. Such a process offers scientists a thermodynamic perspective that is invaluable for material design.

**Quantifying and understanding artifacts in APT Data**: Apart from simplifying assumptions taken in the reconstruction algorithms, artifacts in APT reconstructed data can appear from various

physical phenomena during the experiment. The inhomogeneous distribution of the electric field on the surface of the APT specimen can cause aberrations in the trajectory of the field-evaporated ions [14], [30]. This can lead to the formation of low-density regions on the detector, also called poles, as discussed in Chapter 1. Deviation in ion trajectories can also come from precipitates with different evaporation fields than the surrounding matrix [133]. This can result in erroneous morphology, size, and composition of such precipitates in the reconstructed data [134]. Additionally, atoms that are surrounded by neighbors with different evaporation fields may lead to early or delayed evaporation, causing inaccuracies in the depth coordinates of such atoms [67]. The trajectory aberrations discussed here have an adverse impact on the spatial resolution and local composition [79]. Atomistic simulations can be performed to quantify and understand the resulting artifacts. However, in the current state, such simulations are time-consuming for dataset sizes comparable to the experiments. This can be accelerated by developing accurate machine learning potentials.

**Machine learning based reconstruction algorithms**: Since the detector coordinates of evaporated atoms in atomistic simulations can be deduced, an ML model can be used to learn the correlation between the original location of the atom in the simulated specimen and the detector coordinates. Such a model can then be used to accurately reconstruct experimental data of the same material.

**Creating APT datasets using inverse modeling** Inverse modeling here means predicting the microstructure of a certain material for a desired property [135]. Generating APT datasets through experiments is both time-intensive and expensive, requiring specialized atom probe equipment. Instead, existing experimental data can be used while employing generative ML models to produce new 3D APT datasets at different processing conditions and material properties. This process, however, depends on having a substantial amount of labeled training data. As an example, consider using a variational autoencoder network (VAE), similar to the autoencoder architecture I discussed earlier. We can train this network on the spatial arrangement of voxel centroids. These centroids represent different phases and underlying microstructure and are obtainable after implementing clustering in composition space-based approach on APT datasets. Once the VAE is trained, its encoder segment effectively compresses all crucial information into a latent space. From here, we can sample the latent space according to the required property to generate novel 3D microstructures represented by voxel centroids. This approach is similar to the VAE-based inverse model proposed by Kusampudi et al. for designing 2D microstructures with varied properties, where the training data was curated using crystal plasticity simulations [136].

**Predicting structure-property relationships using APT data**: One of the aims of this thesis was to identify descriptors within APT datasets that effectively represent the underlying information and serve as input to high throughput ML models for finding structure-property relationships. However, two key challenges must be solved to achieve the latter part of this goal. The first is collecting enough APT data with processing conditions and property labels for a given material. This challenge is gradually being overcome as more research groups adopt FAIR (Findable, Accessible, Interoperable, and Reusable) principles [137], which aid in data management and distribution. On top of this, workflow management tools like Apache Airflow [138] can be used to maintain ML pipelines and automatically supplement the existing data with more training data. The second challenge arises from the limited field of view in APT, which restricts us to only local microstructural information [7]. To counter this, APT-derived descriptors can be augmented with data from other experimental and simulation methods to add information at different length scales. These can include SEM (Scanning Electron Microscopy) images, scanning or transmission electron microscopy images, thermodynamic databases (for instance, using CALPHAD - Calculation of Phase Diagrams), and data from simulation techniques like phase field modeling or DFT (Density Functional Theory) [139]–[142]. This approach

is inspired by the work of Rao et al., who successfully enhanced the accuracy of ML models by integrating simulation-based descriptors with experimental data [143].

In summary, this thesis demonstrates the enormous potential of ML-based approaches for APT analysis. My work has accomplished various milestones but has probably only scratched the surface of the hidden data treasure lying in APT datasets.

# Copyright

This thesis includes figures reproduced with permission from publishers. Proper attribution is provided below:

- **Figure 1.1(A):** Reproduced with permission from [7]. Copyright © (2021) Springer Nature.
- **Figure 1.1(B):** Reproduced with permission from [8]. Copyright © (2012) Elsevier.
- **Figure 1.2:** Reproduced with permission from [7]. Copyright © (2021) Springer Nature.

# Bibliography


[1] Dierk Raabe, Jaber Rezaei Mianroodi, and Jörg Neugebauer. "Accelerating the design of compositionally complex materials via physics-informed artificial intelligence". In: *Nature Computational Science* 3.3 (2023), pp. 198–209. DOI: 10.1038/s43588-023-00412-7.

[2] Dierk Raabe. "The Materials Science behind Sustainable Metals and Alloys". In: *Chemical Reviews* 123.5 (Mar. 2023), pp. 2436–2608. DOI: 10.1021/acs.chemrev.2c00799.

[3] M.R. Gilbert, K. Arakawa, Z. Bergstrom, M.J. Caturla, S.L. Dudarev, F. Gao, A.M. Goryaeva, S.Y. Hu, X. Hu, R.J. Kurtz, A. Litnovsky, J. Marian, M.-C. Marinica, E. Martinez, E.A. Marquis, D.R. Mason, B.N. Nguyen, P. Olsson, Y. Osetskiy, D. Senor, W. Setyawan, M.P. Short, T. Suzudo, J.R. Trelewicz, T. Tsuru, G.S. Was, B.D. Wirth, L. Yang, Y. Zhang, and S.J. Zinkle. "Perspectives on multiscale modelling and experiments to accelerate materials development for fusion". In: *Journal of Nuclear Materials* 554 (2021), p. 153113. ISSN: 0022-3115. DOI: 10.1016/j.jnucmat.2021.153113.

[4] Paul F. McMillan. "New materials from high-pressure experiments". In: *Nature Materials* 1.1 (2002), pp. 19–25. DOI: 10.1038/nmat716.

[5] Dongbo Dai, Qing Liu, Rui Hu, Xiao Wei, Guangtai Ding, Baoyu Xu, Tao Xu, Jincang Zhang, Yan Xu, and Huiran Zhang. "Method construction of structure-property relationships from data by machine learning assisted mining for materials design applications". In: *Materials & Design* 196 (2020), p. 109194. ISSN: 0264-1275. DOI: 10.1016/j.matdes.2020.109194.

[6] Huan Zhao, Liam Huber, Wenjun Lu, Nicolas J. Peter, Dayong An, Frédéric De Geuser, Gerhard Dehm, Dirk Ponge, Jörg Neugebauer, Baptiste Gault, and Dierk Raabe. "Interplay of Chemistry and Faceting at Grain Boundaries in a Model Al Alloy". In: *Phys. Rev. Lett.* 124 (10 2020), p. 106102. DOI: 10.1103/PhysRevLett.124.106102.

[7] Baptiste Gault, Ann Chiaramonti, Oana Cojocaru-Mirédin, Patrick Stender, Renelle Dubosq, Christoph Freysoldt, Surendra Kumar Makineni, Tong Li, Michael Moody, and Julie M. Cairney. "Atom probe tomography". In: *Nature Reviews Methods Primers* 1.1 (July 2021). DOI: 10.1038/s43586-021-00047-w.

[8] Baptiste Gault, Michael P. Moody, Julie M. Cairney, and Simon P. Ringer. "Atom probe crystallography". In: *Materials Today* 15.9 (2012), pp. 378–386. ISSN: 1369-7021. DOI: 10.1016/S1369-7021(12)70164-5.

[9] Williams Lefebvre, Francois Vurpillot, and Xavier Sauvage, eds. *Atom probe tomography: Put Theory Into Practice*. San Diego, CA: Academic Press, 2016. URL: https://www.sciencedirect.com/book/9780128046470/atom-probe-tomography.

[10] Michael K. Miller and Richard G. Forbes. *Atom-Probe Tomography*. New York, NY: Springer New York, 2014. DOI: 10.1007/978-1-4899-7430-3.

[11] T. S. Prithiv, Zachary Kloenne, Dian Li, Rongpei Shi, Yufeng Zheng, Hamish L. Fraser, Baptiste Gault, and Stoichko Antonov. "Grain boundary segregation and its implications regarding the formation of the grain boundary $\alpha$ phase in the metastable $\beta$-Titanium Ti–5Al–5Mo–5V–3Cr alloy". In: *Scripta Materialia* 207 (2022), p. 114320. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2021.114320.

[12] X. Zhou, J. R. Mianroodi, A. Kwiatkowski da Silva, T. Koenig, G. B. Thompson, P. Shanthraj, D. Ponge, B. Gault, B. Svendsen, and D. Raabe. "The hidden structure dependence of the chemical life of dislocations". In: *Science Advances* 7.16 (2021), 1–9. DOI: 10.1126/sciadv.abf0563.

[13] S. Medrano, H. Zhao, F. De Geuser, B. Gault, L.T. Stephenson, A. Deschamps, D. Ponge, D. Raabe, and C.W. Sinclair. "Cluster hardening in Al-3Mg triggered by small Cu additions". In: *Acta Materialia* 161 (2018), pp. 12–20. ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2018.08.050.

[14] A.R. Waugh, E.D. Boyes, and M.J. Southon. "Investigations of field evaporation with a field-desorption microscope". In: *Surface Science* 61.1 (1976), pp. 109–142. ISSN: 0039-6028. DOI: https://doi.org/10.1016/0039-6028(76)90411-8.

[15] B. Gault, D. Haley, F. de Geuser, M.P. Moody, E.A. Marquis, D.J. Larson, and B.P. Geiser. "Advances in the reconstruction of atom probe tomography data". In: *Ultramicroscopy* 111.6 (2011). Special Issue: 52nd International Field Emission Symposium, pp. 448–457. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2010.11.016.

[16] Thomas F. Kelly, Tye T. Gribb, Jesse D. Olson, Richard L. Martens, Jeffrey D. Shepard, Scott A. Wiener, Thomas C. Kunicki, Robert M. Ulfig, Daniel R. Lenz, Eric M. Strennen, Edward Oltman, Joseph H. Bunton, and David R. Strait. "First Data from a Commercial Local Electrode Atom Probe (LEAP)". In: *Microscopy and Microanalysis* 10.3 (June 2004), pp. 373–383. ISSN: 1431-9276. DOI: 10.1017/S1431927604040565.

[17] B. Gault, F. Vurpillot, A. Vella, M. Gilbert, A. Menand, D. Blavette, and B. Deconihout. "Design of a femtosecond laser assisted tomographic atom probe". In: *Review of Scientific Instruments* 77.4 (2006), p. 43705. ISSN: 00346748. DOI: 10.1063/1.2194089.

[18] A. Cerezo, P. H. Clifton, A. Gomberg, and G. D. W. Smith. "Aspects of the performance of a femtosecond laser-pulsed 3-dimensional atom probe." In: *Ultramicroscopy* 107.9 (2007), pp. 720–725. ISSN: 0304-3991. DOI: 10.1016/j.ultramic.2007.02.025.

[19] D. Hudson, G.D.W. Smith, and B. Gault. "Optimisation of mass ranging for atom probe microanalysis and application to the corrosion processes in Zr alloys". In: *Ultramicroscopy* 111.6 (2011). Special Issue: 52nd International Field Emission Symposium, pp. 480–486. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2010.11.007.

[20] Olof C. Hellman, John Blatz du Rivage, and David N. Seidman. "Efficient sampling for three-dimensional atom probe microscopy data". In: *Ultramicroscopy* 95 (2003). IFES 2001, pp. 199–205. ISSN: 0304-3991. DOI: https://doi.org/10.1016/S0304-3991(02)00317-0.

[21] Olof C. Hellman, Justin A. Vandenbroucke, Järg Rüsing, Dieter Isheim, and David N. Seidman. "Analysis of Three-dimensional Atom-probe Data by the Proximity Histogram". In: *Microscopy and Microanalysis* 6.5 (2000), 437–444. DOI: 10.1007/S100050010051.

[22] Peter Felfer and Julie Cairney. "Advanced concentration analysis of atom probe tomography data: Local proximity histograms and pseudo-2D concentration maps". In: *Ultramicroscopy* 189 (2018), pp. 61–64. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2018.03.011.

[23] T. Philippe, F. De Geuser, S. Duguay, W. Lefebvre, O. Cojocaru-Mirédin, G. Da Costa, and D. Blavette. "Clustering and nearest neighbour distances in atom-probe tomography". In: *Ultramicroscopy* 109.10 (2009), pp. 1304–1309. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2009.06.007.

[24] M. K. Miller and E. A. Kenik. "Atom Probe Tomography: A Technique for Nanoscale Characterization". In: *Microscopy and Microanalysis* 8.S02 (2002), 1126–1127. DOI: https://doi.org/10.1017/S1431927602103709.

[25] Richard A. Karnesky, Chantal K. Sudbrack, and David N. Seidman. "Best-fit ellipsoids of atom-probe tomographic data to study coalescence of $\gamma'$ (L12) precipitates in Ni–Al–Cr". In: *Scripta Materialia* 57.4 (2007), pp. 353–356. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2007.04.020.

[26] R.K.W. Marceau, L.T. Stephenson, C.R. Hutchinson, and S.P. Ringer. "Quantitative atom probe analysis of nanostructure containing clusters and precipitates with multiple length scales". In: *Ultramicroscopy* 111.6 (2011). Special Issue: 52nd International Field Emission Symposium, pp. 738–742. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2010.12.029.

[27] Leigh T. Stephenson, Michael P. Moody, Peter V. Liddicoat, and Simon P. Ringer. "New techniques for the analysis of fine-scaled clustering phenomena within Atom Probe Tomography (APT) data". In: *Microscopy and Microanalysis* 13.6 (2007), pp. 448–463. ISSN: 14319276. DOI: 10.1017/S1431927607070900.

[28] Jennifer Zelenty, Andrew Dahl, Jonathan Hyde, George D.W. Smith, and Michael P. Moody. "Detecting Clusters in Atom Probe Data with Gaussian Mixture Models". In: *Microscopy and Microanalysis* 23.2 (2017), pp. 269–278. ISSN: 14358115. DOI: 10.1017/S1431927617000320.

[29] A.J.W. Moore. "The structure of atomically smooth spherical surfaces". In: *Journal of Physics and Chemistry of Solids* 23.7 (1962), pp. 907–912. ISSN: 0022-3697. DOI: https://doi.org/10.1016/0022-3697(62)90148-8.

[30] F. Vurpillot, A. Bostel, and D. Blavette. "Trajectory overlaps and local magnification in three-dimensional atom probe". In: *Applied Physics Letters* 76.21 (2000). Cited by: 374, pp. 3127 –3129. DOI: 10.1063/1.126545.

[31] Baptiste Gault, Benjamin Klaes, Felipe F. Morgado, Christoph Freysoldt, Yue Li, Frederic De Geuser, Leigh T. Stephenson, and François Vurpillot. "Reflections on the Spatial Performance of Atom Probe Tomography in the Analysis of Atomic Neighborhoods". In: *Microscopy and Microanalysis* 28.4 (2022), 1116–1126. DOI: 10.1017/S1431927621012952.

[32] Michael P. Moody, Baptiste Gault, Leigh T. Stephenson, Daniel Haley, and Simon P. Ringer. "Qualification of the tomographic reconstruction in atom probe by advanced spatial distribution map techniques". In: *Ultramicroscopy* 109.7 (2009). Cited by: 121, pp. 815 –824. DOI: 10.1016/j.ultramic.2009.03.016.

[33] F. Vurpillot, F. De Geuser, G. Da Costa, and D. Blavette. "Application of Fourier transform and autocorrelation to cluster identification in the three-dimensional atom probe". In: *Journal of Microscopy* 216.3 (2004). Cited by: 54, pp. 234 –240. DOI: 10.1111/j.0022-2720.2004.01413.x.

[34] L. Yao, M.P. Moody, J.M. Cairney, Daniel Haley, A.V. Ceguerra, C. Zhu, and S.P. Ringer. "Crystallographic structural analysis in atom probe microscopy via 3D Hough transformation". In: *Ultramicroscopy* 111.6 (2011). Cited by: 56, pp. 458 –463. DOI: 10.1016/j.ultramic.2010.11.018.

[35] Brian P. Geiser, Thomas F. Kelly, David J. Larson, Jason Schneir, and Jay P. Roberts. "Spatial distribution maps for atom probe tomography". In: *Microscopy and Microanalysis* 13.6 (2007). Cited by: 151, pp. 437 –447. DOI: 10.1017/S1431927607070948.

[36] Baptiste Gault, Frederic de Geuser, Leigh T. Stephenson, Michael P. Moody, Barrington C. Muddle, and Simon P. Ringer. "Estimation of the Reconstruction Parameters for Atom Probe Tomography". In: *Microscopy and Microanalysis* 14.4 (2008), 296–305. DOI: 10.1017/S1431927608080690.

[37] Steven M. Reddy, David W. Saxey, William D. A. Rickard, Denis Fougerouse, Stephanie D. Montalvo, Rick Verberne, and Arie van Riessen. "Atom Probe Tomography: Development and Application to the Geosciences". In: *Geostandards and Geoanalytical Research* 44.1 (Feb. 2020), pp. 5–50. DOI: 10.1111/ggr.12313.

[38] Brian Langelier, Xiaoyue Wang, and Kathryn Grandfield. "Atomic scale chemical tomography of human bone". In: *Scientific Reports* 7.1 (2017), p. 39958. DOI: 10.1038/srep39958.

[39] Michael P. Moody, Fengzai Tang, Baptiste Gault, Simon P. Ringer, and Julie M. Cairney. "Atom probe crystallography: Characterization of grain boundary orientation relationships in nanocrystalline aluminium". In: *Ultramicroscopy* 111.6 (2011). Special Issue: 52nd International Field Emission Symposium, pp. 493–499. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2010.11.014.

[40] Sandeep Madireddy, Ding-Wen Chung, Troy Loeffler, Subramanian K. R. S. Sankaranarayanan, David N. Seidman, Prasanna Balaprakash, and Olle Heinonen. "Phase Segmentation in Atom-Probe Tomography Using Deep Learning-Based Edge Detection". In: *Scientific Reports* 9.1 (2019), p. 20140. ISSN: 2045-2322. DOI: 10.1038/s41598-019-56649-8.

[41] Ye Wei, Baptiste Gault, Rama Srinivas Varanasi, Dierk Raabe, Michael Herbig, and Andrew J. Breen. "Machine-learning-based atom probe crystallographic analysis". In: *Ultramicroscopy* 194 (2018), pp. 15–24. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2018.06.017.

[42] Ye Wei, Rama Srinivas Varanasi, Torsten Schwarz, Leonie Gomell, Huan Zhao, David J. Larson, Binhan Sun, Geng Liu, Hao Chen, Dierk Raabe, and Baptiste Gault. "Machine-learning-enhanced time-of-flight mass spectrometry analysis". In: *Patterns* 2.2 (2021), p. 100192. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2020.100192.

[43] Yue Li, Ye Wei, Zhangwei Wang, Timoteo Colnaghi, Liuliu Han, Ziyuan Rao, Xuyang Zhou, Liam Huber, Raynol Dsouza, Jörg Neugebauer, Andreas Marek, Markus Rampp, Stefan Bauer, Hongxiang Li, Ian Baker, Leigh Stephenson, and Baptiste Gault. *Quantitative three-dimensional imaging of chemical short-range order via machine learning enhanced atom probe tomography*. 2022. DOI: 10.21203/rs.3.rs-1092384/v1.

[44] Yue Li, Xuyang Zhou, Timoteo Colnaghi, Ye Wei, Andreas Marek, Hongxiang Li, Stefan Bauer, Markus Rampp, and Leigh T. Stephenson. "Convolutional neural network-assisted recognition of nanoscale L12 ordered structures in face-centred cubic alloys". In: *npj Computational Materials* 7.1 (2021), p. 8. DOI: 10.1038/s41524-020-00472-7.

[45] Ye Wei, Zirong Peng, Markus Kühbach, Andrew Breen, Marc Legros, Melvyn Larranaga, Frederic Mompiou, and Baptiste Gault. "3D nanostructural characterisation of grain boundaries in atom probe data utilising machine learning methods". In: *PLOS ONE* 14.11 (Nov. 2019). Ed. by Arun Devaraj, e0225041. DOI: 10.1371/journal.pone.0225041.

[46] Xuyang Zhou, Ye Wei, Markus Kühbach, Huan Zhao, Florian Vogel, Reza Darvishi Kamachali, Gregory B. Thompson, Dierk Raabe, and Baptiste Gault. "Revealing in-plane grain boundary composition features through machine learning from atom probe tomography data". In: *Acta Materialia* 226.January (2022), p. 117633. ISSN: 13596454. DOI: 10.1016/j.actamat.2022.117633.

[47] Jan Janssen, Sudarsan Surendralal, Yury Lysogorskiy, Mira Todorova, Tilmann Hickel, Ralf Drautz, and Jörg Neugebauer. "pyiron: An integrated development environment for computational materials science". In: *Computational Materials Science* 163 (2019), pp. 24–36. ISSN: 0927-0256. DOI: https://doi.org/10.1016/j.commatsci.2018.07.043.

[48] Alaukik Saxena, Nikita Polin, Navyanth Kusampudi, Shyam Katnagallu, Leopoldo Molina-Luna, Oliver Gutfleisch, Benjamin Berkels, Baptiste Gault, Jörg Neugebauer, and Christoph Freysoldt. "A Machine Learning Framework for Quantifying Chemical Segregation and Microstructural Features in Atom Probe Tomography Data". In: *Microscopy and Microanalysis* 29.5 (Aug. 2023), pp. 1658–1670. ISSN: 1431-9276. DOI: 10.1093/micmic/ozad086.

[49] Alaukik Saxena, Markus Kühbach, Shyam Katnagallu, Paraskevas Kontis, Baptiste Gault, and Christoph Freysoldt. "Analyzing Linear Features in Atom Probe Tomography Datasets using Skeletonization". In: *Microscopy and Microanalysis* 30.Supplement_1 (July 2024), ozae044.022. ISSN: 1431-9276. DOI: 10.1093/mam/ozae044.022.

[50] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: https://link.springer.com/book/10.1007/978-0-387-84858-7.

[51] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[52] Tom M. Mitchell. *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997. URL: https://www.cs.cmu.edu/~tom/mlbook.html.

[53] Aurelien Geron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. O'Reilly Media, Inc., 2019. ISBN: 1492032646. URL: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

[54] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[55] J. R. Quinlan. "Induction of Decision Trees". In: *Machine Learning* 1.1 (1986), pp. 81–106. URL: https://hunch.net/~coms-4771/quinlan.pdf.

[56] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, 807–814. ISBN: 9781605589077. URL: https://icml.cc/Conferences/2010/papers/432.pdf.

[57] George Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of Control, Signals and Systems* 2 (1989), pp. 303–314. URL: https://api.semanticscholar.org/CorpusID:3958369.

[58] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. "Efficient BackProp". In: *Neural Networks: Tricks of the Trade: Second Edition*. Ed. by Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48. ISBN: 978-3-642-35289-8. DOI: 10.1007/978-3-642-35289-8_3.

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Cham: Springer International Publishing, 2015, pp. 234–241. ISBN: 978-3-319-24574-4. URL: https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28.

[60] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. URL: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

[61] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38. ISSN: 00359246. URL: http://www.jstor.org/stable/2984875.

[62] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, 1996, pp. 226–231. URL: https://www.osti.gov/biblio/421283.

[63] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[64] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. "UMAP: Uniform Manifold Approximation and Projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861.

[65] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[66] D. Vaumousse, A. Cerezo, and P.J. Warren. "A procedure for quantification of precipitate microstructures from three-dimensional atom probe data". In: *Ultramicroscopy* 95 (2003). IFES 2001, pp. 215–221. ISSN: 0304-3991. DOI: https://doi.org/10.1016/S0304-3991(02)00319-4.

[67] Emmanuelle A. Marquis and Francois Vurpillot. "Chromatic Aberrations in the Field Evaporation Behavior of Small Precipitates". In: *Microscopy and Microanalysis* 14.6 (Nov. 2008), pp. 561–570. ISSN: 1431-9276. DOI: 10.1017/S1431927608080793.

[68] William E. Lorensen and Harvey E. Cline. "Marching Cubes: A High Resolution 3D Surface Construction Algorithm". In: *SIGGRAPH Comput. Graph.* 21.4 (1987), pp. 163–169. ISSN: 0097-8930. DOI: 10.1145/37402.37422.

[69] F. Exertier, A. La Fontaine, C. Corcoran, S. Piazolo, E. Belousova, Z. Peng, B. Gault, D.W. Saxey, D. Fougerouse, S.M. Reddy, S. Pedrazzini, P.A.J. Bagot, M.P. Moody, B. Langelier, D.E. Moser, G.A. Botton, F. Vogel, G.B. Thompson, P.T. Blanchard, A.N. Chiaramonti, D.A. Reinhard, K.P. Rice, D.K. Schreiber, K. Kruska, J. Wang, and J.M. Cairney. "Atom probe tomography analysis of the reference zircon gj-1: An interlaboratory study". In: *Chemical Geology* 495 (2018), pp. 27–35. ISSN: 0009-2541. DOI: https://doi.org/10.1016/j.chemgeo.2018.07.031.

[70] Dallin J. Barton, B. Chad Hornbuckle, Kristopher A. Darling, and Gregory B. Thompson. "The Influence of Isoconcentration Surface Selection in Quantitative Outputs from Proximity Histograms". In: *Microscopy and Microanalysis* 25.2 (2019), pp. 401–409. DOI: 10.1017/S143192761900014X.

[71] Yan Dong, Auriane Etienne, Alex Frolov, Svetlana Fedotova, Katsuhiko Fujii, Koji Fukuya, Constantinos Hatzoglou, Evgenia Kuleshova, Kristina Lindgren, Andrew London, and et al. "Atom Probe Tomography Interlaboratory Study on Clustering Analysis in Experimental Data Using the Maximum Separation Distance Approach". In: *Microscopy and Microanalysis* 25.2 (2019), pp. 356–366. DOI: 10.1017/S1431927618015581.

[72] I. Ghamarian and E.A. Marquis. "Hierarchical density-based cluster analysis framework for atom probe tomography data". In: *Ultramicroscopy* 200 (2019), pp. 28–38. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2019.01.011.

[73] Jing Wang, Daniel K. Schreiber, Nathan Bailey, Peter Hosemann, and Mychailo B. Toloczko. "The Application of the OPTICS Algorithm to Cluster Analysis in Atom Probe Tomography Data". In: *Microscopy and Microanalysis* 25.2 (2019), pp. 338–348. ISSN: 14358115. DOI: 10.1017/S1431927618015386.

[74] M. Duerrschnabel, M. Yi, K. Uestuener, M. Liesegang, M. Katter, H. J. Kleebe, B. Xu, O. Gutfleisch, and L. Molina-Luna. "Atomic structure and domain wall pinning in samarium-cobalt-based permanent magnets". In: *Nature Communications* 8.1 (July 2017), pp. 1–7. ISSN: 20411723. DOI: 10.1038/s41467-017-00059-9.

[75] K. Thompson, D. Lawrence, D.J. Larson, J.D. Olson, T.F. Kelly, and B. Gorman. "In situ site-specific specimen preparation for atom probe tomography". In: *Ultramicroscopy* 107.2 (2007), pp. 131–139. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2006.06.008.

[76] Aparna Saksena, Dorota Kubacka, Baptiste Gault, Erdmann Spiecker, and Paraskevas Kontis. "The effect of $\gamma$ matrix channel width on the compositional evolution in a multi-component nickel-based superalloy". In: *Scripta Materialia* 219 (2022), p. 114853. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2022.114853.

[77] K.L. Torres, M. Daniil, M.A. Willard, and G.B. Thompson. "The influence of voxel size on atom probe tomography data". In: *Ultramicroscopy* 111.6 (2011). Special Issue: 52nd International Field Emission Symposium, pp. 464–468. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2011.01.001.

[78] Daniel R. Mason and Andrew J. London. "Morphological analysis of 3d atom probe data using Minkowski functionals". In: *Ultramicroscopy* 211 (2020), p. 112940. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2020.112940.

[79] Baptiste Gault, Michael P. Moody, Julie M. Cairney, and Simon P. Ringer. *Atom Probe Microscopy*. New York, NY: Springer New York, 2012. DOI: 10.1007/978-1-4614-3436-8.

[80] F. Vurpillot, A. Bostel, and D. Blavette. "Trajectory overlaps and local magnification in three-dimensional atom probe". In: *Applied Physics Letters* 76.21 (2000), pp. 3127–3129. DOI: 10.1063/1.126545.

[81] 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 E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[82] Markus Kühbach, Vitor Vieira Rielli, Sophie Primig, Alaukik Saxena, David Mayweg, Benjamin Jenkins, Stoichko Antonov, Alexander Reichmann, Stefan Kardos, Lorenz Romaner, and Sandor Brockhauser. *On Strong-Scaling and Open-Source Tools for High-Throughput Quantification of Material Point Cloud Data: Composition Gradients, Microstructural Object Reconstruction, and Spatial Correlations*. 2022. URL: https://arxiv.org/abs/2205.13510.

[83] Pedro Domingos. "A Few Useful Things to Know about Machine Learning". In: *Commun. ACM* 55.10 (2012), 78–87. ISSN: 0001-0782. DOI: 10.1145/2347736.2347755.

[84] M. Herbig, D. Raabe, Y. J. Li, P. Choi, S. Zaefferer, and S. Goto. "Atomic-Scale Quantification of Grain Boundary Segregation in Nanocrystalline Material". In: *Phys. Rev. Lett.* 112 (12 2014), p. 126103. DOI: 10.1103/PhysRevLett.112.126103.

[85] Zirong Peng, Yifeng Lu, Constantinos Hatzoglou, Alisson Kwiatkowski Da Silva, Francois Vurpillot, Dirk Ponge, Dierk Raabe, and Baptiste Gault. "An Automated Computational Approach for Complete In-Plane Compositional Interface Analysis by Atom Probe Tomography". In: *Microscopy and Microanalysis* 25.2 (2019), pp. 389–400. ISSN: 14358115. DOI: 10.1017/S1431927618016112.

[86] S. Giron, N. Polin, E. Adabifiroozjaei, Y. Yang, A. Kovács, T. P. Almeida, D. Ohmer, K. Üstüner, M. Katter, I. A. Radulov, R. E. Dunin-Borkowski, M. Farle, K. Durst, H. Zhang, L. Alff, K. Ollefs, B. X. Xu, O. Gutfleisch, L. Molina-Luna, B. Gault, and K. P. Skokov. *Towards engineering the perfect defect in high-performing permanent magnets*. 2023. URL: https://arxiv.org/abs/2304.14958.

[87] A.C. Day, H. Francois-Saint-Cyr, B.P Geiser, T. Payne, E. Oltman, S.P. Ringer, and D.A. Reinhard. "Recent Developments in APT Analysis Automation and Support for User-Defined Custom Analysis Procedures in IVAS 4". In: *Microscopy and Microanalysis* 25.S2 (2019), pp. 338–339. DOI: 10.1017/S1431927619002423.

[88] The HDF Group. *Hierarchical Data Format, version 5*. 1997-2022. URL: https://www.hdfgroup.org/HDF5/.

[89] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87 –90. DOI: 10.3233/978-1-61499-649-1-87.

[90] James P. Ahrens, Berk Geveci, and C. Charles Law. "36 - ParaView: An End-User Tool for Large-Data Visualization". In: *Visualization Handbook*. Ed. by Charles D. Hansen and Chris R. Johnson. Burlington: Butterworth-Heinemann, 2005, pp. 717–731. ISBN: 978-0-12-387582-2. DOI: https://doi.org/10.1016/B978-012387582-2/50038-1.

[91] Ian Jolliffe. *Principal component analysis*. New York: Springer Verlag, 2002. URL: https://link.springer.com/book/10.1007/b98835.

[92] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R J Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E A Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy

1.0 Contributors. "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. In: *Nat. Methods* 17.3 (2020), pp. 261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2.

[93] M. Katter, J. Weber, W. Assmus, P. Schrey, and W. Rodewald. "A new model for the coercivity mechanism of $Sm_2$(Co,Fe,Cu,Zr)$_{17}$ magnets". In: *IEEE Transactions on Magnetics* 32.5 (1996), pp. 4815–4817. DOI: 10.1109/20.539161.

[94] Muna Khushaim, Torben Boll, Judith Seibert, Ferdinand Haider, and Talaat Al-Kassab. "Characterization of Precipitation in Al-Li Alloy AA2195 by means of Atom Probe Tomography and Transmission Electron Microscopy". In: *Advances in Condensed Matter Physics* 2015 (2015), p. 647468. ISSN: 16878124. DOI: 10.1155/2015/647468.

[95] M. W. Zandbergen, Q. Xu, A. Cerezo, and G. D.W. Smith. "Study of precipitation in Al-Mg-Si alloys by Atom Probe Tomography I. Microstructural changes as a function of ageing temperature". In: *Acta Materialia* 101 (2015), pp. 136–148. ISSN: 13596454. DOI: 10.1016/j.actamat.2015.08.017.

[96] F. Vogel, N. Wanderka, Z. Balogh, M. Ibrahim, P. Stender, G. Schmitz, T. Fedorova, and J. Banhart. "Evolution of nanoscale clusters in $\gamma'$ precipitates of a Ni-Al-Ti model alloy". In: *Ultramicroscopy* 159.April 2021 (2015), pp. 278–284. ISSN: 18792723. DOI: 10.1016/j.ultramic.2015.04.021.

[97] X. Zhou, J. R. Mianroodi, A. Kwiatkowski da Silva, T. Koenig, G. B. Thompson, P. Shanthraj, D. Ponge, B. Gault, B. Svendsen, and D. Raabe. "The hidden structure dependence of the chemical life of dislocations". In: *Science Advances* 7.16 (2021), eabf0563. DOI: 10.1126/sciadv.abf0563.

[98] I. Medouni, A. Portavoce, P. Maugis, P. Eyméoud, M. Yescas, and K. Hoummada. "Role of dislocation elastic field on impurity segregation in Fe-based alloys". In: *Scientific Reports* 11.1 (2021), p. 1780. DOI: 10.1038/s41598-020-80140-4.

[99] A. Prakash, J. Guénolé, J. Wang, J. Müller, E. Spiecker, M.J. Mills, I. Povstugar, P. Choi, D. Raabe, and E. Bitzek. "Atom probe informed simulations of dislocation–precipitate interactions reveal the importance of local interface curvature". In: *Acta Materialia* 92 (2015), pp. 33–45. ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2015.03.050.

[100] A. H. Cottrell and B. A. Bilby. "Dislocation Theory of Yielding and Strain Ageing of Iron". In: *Proceedings of the Physical Society. Section A* 62.1 (1949), p. 49. DOI: 10.1088/0370-1298/62/1/308.

[101] M. Kuzmina, M. Herbig, D. Ponge, S. Sandlöbes, and D. Raabe. "Linear complexions: Confined chemical and structural states at dislocations". In: *Science* 349.6252 (2015), pp. 1080–1083. DOI: 10.1126/science.aab2633.

[102] Iman Ghamarian, Li-Jen Yu, and Emmanuelle A. Marquis. "Quantification of Solute Topology in Atom Probe Tomography Data: Application to the Microstructure of a Proton-Irradiated Alloy 625". In: *Metallurgical and Materials Transactions A* 51.1 (2020), pp. 42–50. DOI: 10.1007/s11661-019-05520-6.

[103] Peter Felfer, Anna Ceguerra, Simon Ringer, and Julie Cairney. "Applying computational geometry techniques for advanced feature analysis in atom probe data". In: *Ultramicroscopy* 132 (2013). IFES 2012, pp. 100–106. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2013.03.004.

[104] J.M. Hyde, M.G. Burke, B. Gault, D.Wf. Saxey, P. Styman, K.B. Wilford, and T.J. Williams. "Atom probe tomography of reactor pressure vessel steels: An analysis of data integrity". In: *Ultramicroscopy* 111.6 (2011). Special Issue: 52nd International Field Emission Symposium, pp. 676–682. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2010.12.033.

[105] I. Ghamarian, L.-J. Yu, and E.A. Marquis. "Morphological classification of dense objects in atom probe tomography data". In: *Ultramicroscopy* 215 (2020), p. 112996. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2020.112996.

[106] Junjie Cao, Andrea Tagliasacchi, Matt Olson, Hao Zhang, and Zhixun Su. "Point Cloud Skeletons via Laplacian Based Contraction". In: *2010 Shape Modeling International Conference* (2010), pp. 187–197. DOI: 10.1109/SMI.2010.25.

[107] Andrea Tagliasacchi, Ibraheem Alhashim, Matt Olson, and Hao Zhang. "Mean Curvature Skeletons". In: *Computer Graphics Forum* 31.5 (2012), pp. 1735–1744. DOI: https://doi.org/10.1111/j.1467-8659.2012.03178.x.

[108] A. Cervellon, S. Hémery, P. Kürnsteiner, B. Gault, P. Kontis, and J. Cormier. "Crack initiation mechanisms during very high cycle fatigue of Ni-based single crystal superalloys at high temperature". In: *Acta Materialia* 188 (2020), pp. 131–144. ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2020.02.012.

[109] A. Kwiatkowski da Silva, D. Ponge, Z. Peng, G. Inden, Y. Lu, A. Breen, B. Gault, and D. Raabe. "Phase nucleation through confined spinodal fluctuations at crystal defects evidenced in Fe-Mn alloys". In: *Nature Communications* 9.1 (2018), p. 1137. DOI: 10.1038/s41467-018-03591-4.

[110] The CGAL Project. *CGAL User and Reference Manual*. 5.6. CGAL Editorial Board, 2023. URL: https://doc.cgal.org/5.6/Manual/packages.html.

[111] Oscar Kin-Chung Au, Chiew-Lan Tai, Hung-Kuo Chu, Daniel Cohen-Or, and Tong-Yee Lee. "Skeleton extraction by mesh contraction". English. In: *ACM Transactions on Graphics* 27.3 (Aug. 2008). ISSN: 0730-0301. DOI: 10.1145/1360612.1360643.

[112] Mathieu Desbrun, Mark Meyer, Peter Schroder, and Alan H. Barr. "Implicit Fairing of Irregular Meshes Using Diffusion and Curvature Flow". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 9798400708978. DOI: https://doi.org/10.1145/3596711.3596729.

[113] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[114] J. A. Nelder and R. Mead. "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313. ISSN: 0010-4620. DOI: 10.1093/comjnl/7.4.308.

[115] Sandra Korte-Kerzel, Tilmann Hickel, Liam Huber, Dierk Raabe, Stefanie Sandlöbes-Haut, Mira Todorova, and Jörg Neugebauer. "Defect phases – thermodynamics and impact on material properties". In: *International Materials Reviews* 67.1 (2022), pp. 89–117. DOI: 10.1080/09506608.2021.1930734.

[116] Shyam Katnagallu, Suzanne Vernier, Marie-Agathe Charpagne, Baptiste Gault, Nathalie Bozzolo, and Paraskevas Kontis. "Nucleation mechanism of hetero-epitaxial recrystallization in wrought nickel-based superalloys". In: *Scripta Materialia* 191 (2021), pp. 7–11. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2020.09.012.

[117] Vicente J. Araullo-Peters, Andrew Breen, Anna V. Ceguerra, Baptiste Gault, Simon P. Ringer, and Julie M. Cairney. "A new systematic framework for crystallographic analysis of atom probe data". In: *Ultramicroscopy* 154 (2015), pp. 7–14. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2015.02.009.

[118] Baptiste Gault, Xiang Yuan Cui, Michael P. Moody, Frederic De Geuser, Christophe Sigli, Simon P. Ringer, and Alexis Deschamps. "Atom probe microscopy investigation of Mg site occupancy within $\delta'$ precipitates in an Al–Mg–Li alloy". In: *Scripta Materialia* 66.11 (2012). Viewpoint set no. 49: Strengthening effect of nano-scale twins, pp. 903–906. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2012.02.021.

[119] Baptiste Gault, Michael P. Moody, Julie M. Cairney, and Simon P. Ringer. "Atom probe crystallography". In: *Materials Today* 15.9 (2012), pp. 378–386. ISSN: 1369-7021. DOI: https://doi.org/10.1016/S1369-7021(12)70164-5.

[120] Ziyuan Rao, Yue Li, Hongbin Zhang, Timoteo Colnaghi, Andreas Marek, Markus Rampp, and Baptiste Gault. "Direct recognition of crystal structures via three-dimensional convolutional neural networks with high accuracy and tolerance to random displacements and missing atoms". In: *Scripta Materialia* 234 (2023), p. 115542. ISSN: 1359-6462. DOI: https://doi.org/10.1016/j.scriptamat.2023.115542.

[121] Albert P. Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments". In: *Phys. Rev. B* 87 (18 2013), p. 184115. DOI: 10.1103/PhysRevB.87.184115.

[122] Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. "Comparing molecules and solids across structural and alchemical space". In: *Phys. Chem. Chem. Phys.* 18 (20 2016), pp. 13754–13769. DOI: 10.1039/C6CP00415F.

[123] Andreas Leitherer, Angelo Ziletti, and Luca M. Ghiringhelli. "Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning". In: *Nature Communications* 12.1 (2021), p. 6234. DOI: 10.1038/s41467-021-26511-5.

[124] Jihan Zhou, Yongsoo Yang, Peter Ercius, and Jianwei Miao. "Atomic electron tomography in three and four dimensions". In: *MRS Bulletin* 45.4 (2020), 290–297. DOI: 10.1557/mrs.2020.88.

[125] Lauri Himanen, Marc O. J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. "DScribe: Library of descriptors for machine learning in materials science". In: *Computer Physics Communications* 247 (2020), p. 106949. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2019.106949.

[126] Jarno Laakso, Lauri Himanen, Henrietta Homm, Eiaki V Morooka, Marc OJ Jäger, Milica Todorović, and Patrick Rinke. "Updates to the DScribe library: New descriptors and derivatives". In: *The Journal of Chemical Physics* 158.23 (2023). DOI: 10.1063/5.0151031.

[127] Jacopo Baima, Alexandra M. Goryaeva, Thomas D. Swinburne, Jean-Bernard Maillet, Maylise Nastar, and Mihai-Cosmin Marinica. "Capabilities and limits of autoencoders for extracting collective variables in atomistic materials science". In: *Phys. Chem. Chem. Phys.* 24 (38 2022), pp. 23152–23163. DOI: 10.1039/D2CP01917E.

[128] Addis S. Fuhr and Bobby G. Sumpter. "Deep Generative Models for Materials Discovery and Machine Learning-Accelerated Innovation". In: *Frontiers in Materials* 9 (2022). ISSN: 2296-8016. DOI: 10.3389/fmats.2022.865270.

[129] Li Deng. "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142. DOI: 10.1109/MSP.2012.2211477.

[130] Renelle Dubosq, Anna Rogowitz, Kevin Schweinar, Baptiste Gault, and David A. Schneider. "A 2D and 3D nanostructural study of naturally deformed pyrite: assessing the links between trace element mobility and defect structures". In: *Contributions to Mineralogy and Petrology* 174.9 (2019), p. 72. DOI: 10.1007/s00410-019-1611-5.

[131] Yarin Gal and Zoubin Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 1050–1059. URL: https://proceedings.mlr.press/v48/gal16.html.

[132] Christian Oberdorfer, Sebastian Manuel Eich, and Guido Schmitz. "A full-scale simulation approach for atom probe tomography". In: *Ultramicroscopy* 128 (2013), pp. 55–67. ISSN: 0304-3991. DOI: https://doi.org/10.1016/j.ultramic.2013.01.005.

[133] F. Vurpillot, A. Bostel, and D. Blavette. "Trajectory overlaps and local magnification in three-dimensional atom probe". In: *Applied Physics Letters* 76.21 (May 2000), pp. 3127–3129. ISSN: 0003-6951. DOI: 10.1063/1.126545.

[134] C. Hatzoglou, B. Radiguet, and P. Pareige. "Experimental artefacts occurring during atom probe tomography analysis of oxide nanoparticles in metallic matrix: Quantification and correction". In: *Journal of Nuclear Materials* 492 (2017), pp. 279–291. ISSN: 0022-3115. DOI: https://doi.org/10.1016/j.jnucmat.2017.05.008.

[135] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. "Inverse molecular design using machine learning: Generative models for matter engineering". In: *Science* 361.6400 (2018), pp. 360–365. DOI: 10.1126/science.aat2663.

[136] Navyanth Kusampudi and Martin Diehl. "Inverse design of dual-phase steel microstructures using generative machine learning model and Bayesian optimization". In: *International Journal of Plasticity* 171 (2023), p. 103776. ISSN: 0749-6419. DOI: https://doi.org/10.1016/j.ijplas.2023.103776.

[137] Claudia Draxl and Matthias Scheffler. "NOMAD: The FAIR concept for big data-driven materials science". In: *MRS Bulletin* 43.9 (2018), 676–682. DOI: 10.1557/mrs.2018.208.

[138] Pramod Singh. "Airflow". In: *Learn PySpark: Build Python-based Machine Learning and Deep Learning Models*. Berkeley, CA: Apress, 2019, pp. 67–84. ISBN: 978-1-4842-4961-1. DOI: 10.1007/978-1-4842-4961-1_4.

[139] Joseph I. Goldstein, Dale E. Newbury, Alton D. Romig, Charles Fiori, and Eric Lifshin. *Scanning Electron Microscopy and X-Ray Microanalysis : a Text for Biologists, Materials Scientists, and Geologists*. N. Boston: Springer US Boston, 1992. ISBN: 9781461304913; 1461304911. DOI: 10.1007/978-1-4613-0491-3.

[140] Hans Lukas, Suzana G. Fries, and Bo Sundman. "Computational Thermodynamics: The Calphad Method". In: Cambridge University Press, 2007. DOI: 10.1017/CBO9780511804137.

[141] Long-Qing Chen. "Phase-Field Models for Microstructure Evolution". In: *Annual Review of Materials Research* 32.1 (2002), pp. 113–140. DOI: 10.1146/annurev.matsci.32.112001.132041.

[142] Nathan Argaman and Guy Makov. "Density functional theory: An introduction". In: *American Journal of Physics* 68.1 (Jan. 2000), pp. 69–79. ISSN: 0002-9505. DOI: 10.1119/1.19375.

[143] Ziyuan Rao, Po-Yen Tung, Ruiwen Xie, Ye Wei, Hongbin Zhang, Alberto Ferrari, T.P.C. Klaver, Fritz Körmann, Prithiv Thoudden Sukumar, Alisson Kwiatkowski da Silva, Yao Chen, Zhiming Li, Dirk Ponge, Jörg Neugebauer, Oliver Gutfleisch, Stefan Bauer, and Dierk Raabe. "Machine learning–enabled high-entropy alloy discovery". In: *Science* 378.6615 (2022), pp. 78–85. DOI: 10.1126/science.abo4940.

# List of Figures

1.1 Schematic of an atom probe ultra-high vacuum (UHV) analysis chamber experimental setup. . . . 2
1.2 Traditional APT data analysis methods. . . . 4

2.1 A schematic of a random forest (RF) model comprising of three decision trees. . . . 9
2.2 A schematic of a decision tree. . . . 10
2.3 A schematic of a deep neural network (DNN). . . . 11
2.4 A schematic depicting the structure of a neuron. . . . 12
2.5 Input and output mask images corresponding to an electron channeling contrast (ECC) micrograph to train a U-Net model. . . . 13
2.6 A schematic of the U-Net model. . . . 13
2.7 A schematic of core, border, and noise points for density-based spatial clustering of applications with noise (DBSCAN) algorithm . . . 16
2.8 Principal components of a two dimensional dataset. . . . 17

3.1 Workflow of clustering in composition space-based approach . . . 21
3.2 Bayesian information criterion (BIC) curve for composition space corresponding to Fe-doped Sm-Co alloy APT dataset . . . 22
3.3 Normalized frequencies of voxel composition (at. fraction) of Cu, Fe, Co, Zr, and Sm for each phase present in one of the Sm-Co alloy APT dataset. . . . 23
3.4 Comparison between isosurfaces in a Fe-doped Sm-Co alloy APT dataset extracted using AP suite and the precipitates extracted at voxel resolution using clustering in composition space. . . . 25
3.5 The dependence of the standard error on subset size (number of voxels) for Z phase in the given Fe-doped Sm-Co alloy. . . . 26
3.6 Comparison between experimental composition distribution and random distribution of 2:17 matrix cell, 1:5 and Z phase precipitates. . . . 27
3.7 APT synthetic data consisting of $L1_2$ precipitate ($Al_3Li$) embedded in an AlLi solid solution. . . . 27
3.8 BIC minimization applied on the composition space of Al-Li synthetic APT data. . . . 28
3.9 Dependence of isosurface volume on the isovalue for the synthetic data. . . . 29
3.10 BIC minimization plot for the composition space for a Ni-based superalloy. . . . 30
3.11 The developed scheme for composition space dimensionality reduction. . . . 30
3.12 Relative importance of various chemical species and BIC minimization in a Ni-based superalloy. . . . 31
3.13 Clusters in composition space of a Ni-based superalloy considering 4 most important features or species . . . 32

4.1 Workflow to quantify a 3D microstructure obtained from an APT reconstructed data. . 36

4.2 Workflow for precipitate disentanglement via junction detection and removal. . . . . 37
4.3 Local PCA and U-Net model based approaches for junction detection. . . . . 38
4.4 The architecture of U-Net image segmentation model used for junction detection. . . . 39
4.5 Comparison between isosurfaces in H3 dataset extracted using AP suite and the precipitates extracted at voxel resolution using clustering in composition space. . . . . 40
4.6 Comparison between isosurfaces in L3 dataset extracted using AP suite and the precipitates extracted at voxel resolution using clustering in composition space. . . . . 41
4.7 Superimposing a triangular grid on a plate-like precipitate. . . . . 43
4.8 Truncated 1D composition profiles across plate-like precipitates in H3 dataset. . . . . 44
4.9 Full 1D composition profiles across plate-like precipitates in H3 dataset. . . . . 45
4.10 In-plane composition fluctuations for precipitates in H3 dataset. . . . . 46
4.11 In-plane thickness fluctuations for precipitates in H3 dataset. . . . . 46
4.12 The number of modes or peaks in the 1D composition profile of Cu in Z phase precipitates of H3 and L3 datasets. . . . . 47
4.13 In-plane thickness fluctuations for Z phase precipitates in H3 and L3 datasets. . . . . 47
4.14 Sets of parallel 1:5 and Z phase precipitates in L2 dataset with distances between them. 48

5.1 Workflow to extract and quantify linear features from APT reconstructed data. . . . . 54
5.2 A B-spline fitted on skeleton nodes. . . . . 56
5.3 Application of mean curvature flow (MCF) algorithm on synthetic APT dataset. . . . . 57
5.4 Application of mean curvature flow (MCF) algorithm on synthetic APT dataset with half-torus morphology. . . . . 57
5.5 The variation of error or the average deviation from the true medial axis of the torus mesh, for each computed MCF skeleton, with values of $W_H$. . . . . 59
5.6 The distribution of the distances from the true medial axis for all the skeleton nodes for the half-torus mesh with $r_{out} = 20$nm and $\mathbf{W}_H = 0.05$. . . . . 60
5.7 Radial fluctuation in the composition for the synthetic APT dataset with half-torus morphology. . . . . 60
5.8 Different ways of placing the ROIs on a half-torus linear feature/dislocation using the fitted B-spline. . . . . 61
5.9 MCF skeleton calculation for Cr isosurface mesh in a Ni-based superalloy APT dataset. 63
5.10 Morphology based segmentation of MCF skeleton. . . . . 64
5.11 Radial fluctuations in the composition of Al, Co, and Cr about the geometric core of a dislocation in a Ni-based superalloy. . . . . 65
5.12 Computing 1D composition profiles across and along a dislocation in a Ni-based superalloy. . . . . 66
5.13 Spherical ROI in an APT dataset with parallel atomic planes. . . . . 68
5.14 MCF skeletons for Mn isosurface meshes in a Fe-Mn alloy APT dataset. . . . . 69
5.15 Poles on a 2D histogram of x-y coordinates of the selected slice of atoms from a Fe-Mn APT dataset. . . . . 69
5.16 Peak strength values calculated after FFT analysis. . . . . 70
5.17 A normal vector to the atomic planes in a Fe-Mn alloy APT dataset. . . . . 70
5.18 The absolute values after applying FFT on the y-coordinate histogram. . . . . 71
5.19 Crystallographic directions extracted after the orientation analysis. . . . . 71

6.1 Workflow for curating synthetic APT data to train a dense neural network for automatic classification of $L1_2$ ordered domains. . . . . 75

6.2 A schematic of autoencoder architecture. . . . . 77
6.3 A schematic of autoencoder architecture with a predictor. . . . . 78
6.4 $L1_2$ and FCC regions in the Al-Mg-Li alloy APT reconstructed data. . . . . 79
6.5 Application of autoencoder with predictor architecture on Al-Mg-Li alloy APT dataset. 80
6.6 Latent space for two different sets of SOAP hyperparameters after training the autoencoder-predictor architecture on the Al-Mg-Li alloy APT dataset. . . . . 81
6.7 Latent space after training an autoencoder-predictor architecture after removing atoms at the edge of the APT dataset. . . . . 82
6.8 Dense neural network architecture trained to get the correlation between the synthetic data SOAP vectors and the corresponding labels. . . . . 83
6.9 Atomic environment predictions of the trained dense neural network. . . . . 83

# List of Tables

3.1 Composition of Ni-based superalloy in at.%. The rest of the amount is Ni. . . . . 21
3.2 Volume fraction, average thickness and average composition of 2:17, Z, 1:5 phase in one of the given APT datasets. . . . . 24

4.1 Average volume fraction, thickness, and composition of each phase of six Fe-doped Sm-Co alloy APT data sets. . . . . 40

5.1 Composition of Ni-based single-crystal superalloy in wt.%. The rest of the amount is Ni. 53
5.2 Composition of Fe-Mn alloy in wt.%. The rest of the amount is Fe. . . . . 53