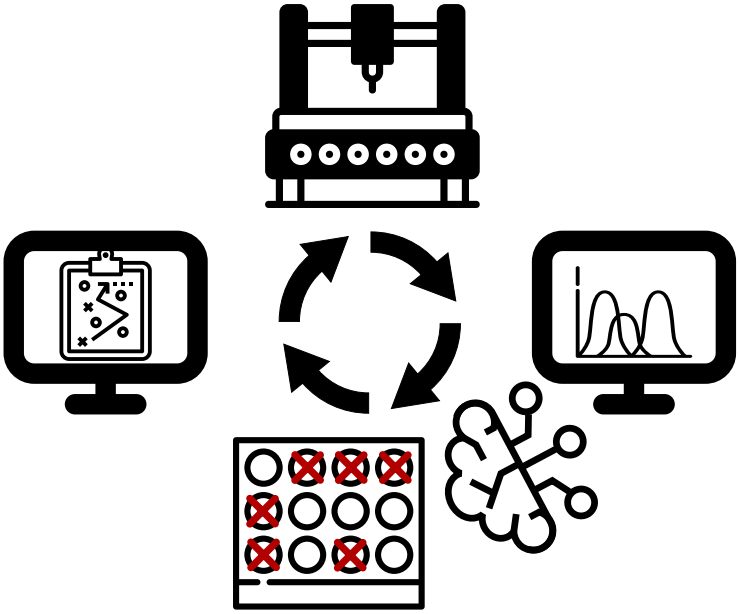
Karlsruher Institut für Technologie

Fakultät für Maschinenbau

Institut für Automation und angewandte Informatik
Maschinelles Lernen für Hochdurchsatzverfahren und Mechatronik

# Spectra-based Neural Networks for Uncovering Novel Substances in Material Discovery Experiments

Jan Schützke



DISSERTATION

# Spectra-based Neural Networks for Uncovering Novel Substances in Material Discovery Experiments

Zur Erlangung des akademischen Grades eines

**DOKTORS DER INGENIEURWISSENSCHAFTEN**
**(Dr.-Ing.)**

von der KIT-Fakultät für Maschinenbau
des Karlsruher Instituts für Technologie (KIT)

angenommene

**DISSERTATION**

von

**M.Sc. Jan Schützke**
geb. in Grünstadt

| | |
|---|---|
| Tag der mündlichen Prüfung: | 17.04.2024 |
| Hauptreferent: | Prof. Dr.-Ing. Markus Reischl |
| Korreferent: | Prof. Dr. Jan G. Korvink |

# Kurzfassung

Analysetechniken wie die Röntgenbeugung (XRD) und die Raman-Spektroskopie sind für die Untersuchung von Materialien, Molekülen und anderen Objekten in Größenordnungen jenseits der menschlichen Wahrnehmung von entscheidender Bedeutung. Diese Methoden sind unverzichtbar, um sowohl bekannte als auch noch nicht entdeckte Stoffe fundamental zu verstehen und zu bewerten. Die im Rahmen der XRD oder Raman-Spektroskopie aufgenommen Daten werden typischerweise als "Spektren" bezeichnet und zeigen einen Intensitätsverlauf in Abhängigkeit von einer variablen Messgröße. In diesen Messungen zeigen die verschiedenen Substanzen einzigartige Muster, ähnlich eines Fingerabdrucks. Dementsprechend lässt sich das Vorhandensein verschiedener Materialien festellen, in dem ihre Fingerabdrücke in den gemessenen Daten nachgewiesen werden.

Da die Nachfrage nach Stoffen mit verbesserten Eigenschaften, z.B. für leistungsfähigere Batterien, Antibiotika ohne Resistenzen oder leichtgewichtige Infrastrukturmaterialien, steigt, liegt der Fokus zunehmend auf der Erforschung von verschiedenen Materialsystemen. Infolgedessen wurden Hochdurchsatzsysteme mit integrierter Robotik entwickelt, um die Herstellung und Untersuchung neuer Materialien und Moleküle zu beschleunigen, welche alle mit Hilfe der oben genannten Techniken analysiert werden. Zur Analyse der Messungen werden jedoch hauptsächlich Methoden eingesetzt, welche manuelle Handhabung erfordern oder das Vorhandensein von Referenzdaten voraussetzen und deswegen die Erforschung der Substanzen oft ausbremsen.

Als Alternative, präsentiert diese Arbeit ein umfangreiches Konzept, welches die Analyse von XRD Mustern und Raman Spektren mittels Einsatz von künstlichen neuronalen Netzwerken automatisiert. Dieses Konzept beinhaltet einen flexiblen

Ansatz zur Simulation von exemplarischen Daten, der die Fingerabdrücke der zu untersuchenden Materialien exakt widerspiegelt. Anschließend wird das neuronale Netz mit Hilfe der simulierten Daten darauf trainiert, die Materialien in den experimentellen Messungen zu erkennen. Die Effektivität dieses Konzepts wird durch seine Anwendung auf drei verschiedene experimentelle Datensätze demonstriert, die jeweils die Herstellung verschiedener Materialien untersucht. Die Ergebnisse zeigen, dass die neuronalen Netze eine schnelle und akkurate Auswertung der gemessenen Signale ermöglichen, obwohl sie auf simulierten Daten trainiert wurden. Darüber hinaus wird die Flexibilität dieses Konzepts hervorgehoben, da es in der Lage ist, Messungen zu analysieren, die mit verschiedenen Messkonfigurationen aufegenommen wurden, ohne, dass eine Anpassung der Methoden notwenig ist. Das entwickelte System lässt sich dementsprechend problemlos in bestehende Hochdurchsatzsysteme integrieren und bietet das Potenzial, die Entdeckung neuer Materialien erheblich zu beschleunigen.

# Abstract

Analytical techniques such as X-ray diffraction (XRD) and Raman spectroscopy are crucial for studying materials, molecules, and other objects at scales beyond human vision. These methods are indispensable for understanding and assessing both known and yet-to-be-discovered substances. They generate one-dimensional intensity patterns, often referred to as "spectra", and each substance has a unique pattern, much like a fingerprint. By examining these patterns, often a combination of fingerprints from known materials and molecules, one can accurately determine a sample's composition. As the demand for substances with enhanced properties, e.g., more efficient batteries, antibiotics resistant to bacterial adaptation, or lightweight infrastructure materials, grows, there is an increasing focus on exploring material compositions to unearth novel discoveries. Consequently, high-throughput systems integrated with robotics have emerged to expedite the production and study of new materials and molecules, all of which are analyzed using the techniques mentioned before. However, the reliance on manual adjustments in traditional methods of analyzing spectra and diffraction patterns frequently becomes a bottleneck in the evaluation process.

Accordingly, a novel framework is introduced that addresses the bottleneck of analyzing XRD patterns and Raman spectra, employing a neural network for automated data analysis. This framework includes a versatile data simulation approach that accurately represents the materials under investigation. Utilizing this synthetic data, the neural network is trained to identify novel materials within experimental signals. The effectiveness of this framework is demonstrated through its application to three distinct experimental datasets, each focused on the formation of different materials. The results highlight that the high predictive quality of the models trained on synthetic data effectively translates to the

analysis of measured signals. Moreover, the flexibility of this framework is emphasized, as it is capable of analyzing scans from various measurement modalities without the need to alter the training data generation or model training methodologies. This developed framework is readily available for integration into existing high-throughput systems, offering the potential to expedite the discovery of new materials significantly.

# Acknowledgements

This thesis marks the end of my journey to obtain a doctoral degree, and I am still in the process of comprehending its conclusion. When I started working as a student researcher at the Institute of Automation and Applied Informatics (IAI) at the Karlsruhe Institute of Technology (KIT) in 2018, I would never have thought that it would turn out the way it did. To this end, I would like to thank Prof. Markus Reischl, who provided me with several interesting projects to explore the field of machine learning and deep learning. I would also like to express gratitude to Prof. Reischl for recommending the pursuit of a doctoral degree and offering a position in his group, a path I had not initially planned during my time as a student. Thanks to his invaluable supervision, I have had many enriching discussions and conversations about my projects and other topics, as well as the financial support that has facilitated the presentation of my research at conferences and during visits to other research groups.

Furthermore, I want to thank Bruker AXS, and, in particular, Dr. Alexander Benedix, for introducing me to the challenges of analyzing X-ray diffraction data. Without the continuous support of my colleagues at Bruker, many of my ideas would have remained unrealized. I would also like to thank the many colleagues in my group at KIT for the discussions about deep learning in general, which allowed me to gain a more general understanding of the neural network models. In particular, my office partner Friedrich provided me with interesting ideas and alternative views to the various challenges, but I also had several interesting conversations with Moritz, Roman, Yanke, Marcel, Luca, and André that resulted in the approaches and methods presented in this thesis. In addition, I would like to thank the other colleagues at IAI who have supported me along the way, including Bernadette, Ralf, Tim, Andy, Katharina, Kaleb, Lorenz, and Oli.

Beyond the IAI, I would like to express my gratitude to individuals and organizations that have played a significant role in this journey. Special thanks to Ben and Simon from the Institute of Nanotechnology at KIT for their help in obtaining the XRD measurements, a task I could not do on my own. Thanks also to Prof. Korvink for reviewing my thesis. I would further like to thank Prof. Ceder, and especially Nathan and Yan, for inviting me to Berkeley and for the many helpful conversations about using deep learning for automated analysis of materials science data. Thanks also to the Karlsruhe House of Young Scientists (KHYS) for supporting my stay in Berkeley with the Research Travel Grant.

Concludingly, I would like to thank my family. I am deeply grateful to my parents for providing everything I needed during my time in Karlsruhe and my travels abroad. And, above all, I am grateful for the invaluable support of my wife Julia throughout this transformative journey.

Karlsruhe, January 2024 *Jan Schützke*

# Contents

# Acronyms and symbols

## Abbreviations

| Abbreviation | Description |
|---|---|
| 1D | One-dimensional |
| CIF | Crystallographic Information File |
| CNN | Convolutional Neural Network |
| COD | Crystallography Open Database |
| DFT | Density Functional Theory |
| DRX | Disordered Rocksalt |
| FoM | Figure-of-Merit |
| FWHM | Full Width Half Maximum |
| GoF | Goodness-of-Fit |
| GNN | Graph Neural Network |
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| GUI | Graphical User Interface |
| IAI | Institute for Automation and Applied Informatics |
| ICSD | Inorganic Crystal Structure Database |
| INT | Institute of Nanotechnology |
| KIT | Karlsruhe Institute of Technology |
| MP | Materials Project |
| MSE | Mean Squared Error |
| NMR | Nuclear Magnetic Resonance |
| OQMD | Open Quantum Materials Database |
| ReLU | Rectified Linear Unit |
| RMSE | Root Mean Squared Error |
| SNR | Signal-to-Noise Ratio |
| XRD | X-ray Diffraction |
| YBCO | Yttrium Barium Copper Oxide |

## Constants

| | |
|---|---|
| $\pi$ | Pi: $3.14159\ldots$ |
| $e$ | Euler's number: $2.71828\ldots$ |
| $K$ | Scherrer constant: $0.9$ |
| **Å** | Angstrom: $10^{-10}$ m |

## Latin symbols and variables

| | |
|---|---|
| $\mathbf{a}, \mathbf{b}, \mathbf{c}$ | Lattice vectors |
| $a, b, c$ | Lattice parameters |
| $h, k, l$ | Miller indices |
| $d_{hkl}$ | Distance of Miller planes |
| $L_{\mathrm{crystallite}}$ | Crystallite size |
| $c_n, o_n$ | Neural Network input/output |
| $w_n$ | Weights for layer in Neural Network |
| $b, \sigma$ | Bias, Activation function in Neural Network |
| $G(x), L(x)$ | Gaussian/ Cauchy probability density function |
| $T_n(x)$ | Chebyshev Polynomial |
| $P(x)$ | Background Function |
| $I_{\mathrm{Signal}}, I_{\mathrm{Noise}}$ | Intensity of Signal, Noise in scan |
| $str_x$ | Crystalline Structure (with lattice, coordinates) of x |
| **A,B,C,D** | End members in phase diagrams |

| | |
|---|---|
| $L, \ell$ | Loss function |

## Greek symbols and variables

| | |
|---|---|
| $\lambda$ | wavelength |
| $\alpha, \beta, \gamma$ | Lattice parameters |
| $\theta, 2\theta$ | Diffraction angle (Bragg's angle) |
| $2\theta_{\min}, 2\theta_{\max}$ | Start, End point range for XRD scan |
| $\Delta 2\theta$ | Step width diffraction angle |
| $K\alpha_1, K\alpha_2, K_\beta$ | Emission lines X-rays |
| $\delta$ | impulse/ Dirac function |
| $\mu, \sigma$ | Mean, standard deviation |
| $\Gamma$ | Cauchy scale parameter |
| $\Phi$ | Paramters of presented neural network |

# 1 Introduction

## 1.1 Motivation

The discovery of novel materials is the key to advancing technological applications in fields such as battery technology. With the increasing demands of energy consumption and storage, it is necessary to develop materials with improved properties that are fit for future conditions [1]. Similarly, there is a need for high-performance materials that can be used to improve characteristics in rocket nozzles [2] or lightweight but strong materials in transport or infrastructure technology [3]. However, the process of discovering new materials through experimental testing is a time-consuming and laborious task.

The conventional process of materials discovery typically involves several steps [4]. Researchers start by mixing precursor materials and synthesizing a few samples with unique compositions identified as promising candidates based on earlier experiments. Those materials are then evaluated using different characterization techniques to determine the relevant properties of the produced samples. For example, the analysis of stresses and strains in materials subjected to external forces is a relevant property for substances in mechanical or aerospace applications. Thus, specimens of these novel materials are prepared for tensile testing, and scientists then interpret the stress-strain curve of the respective samples to determine attributes, such as the ultimate strength or Young's modulus [5]. Alternatively, the synthesized materials are examined via other characterization techniques to determine properties on the atomic scale, conductivity, or magnetic behavior, among many other attributes. These techniques generate large amounts of data that experts in the respective fields need to analyze and interpret [6, 7].

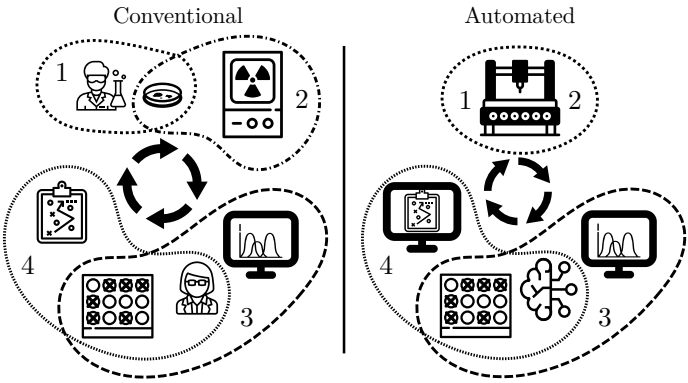



**Figure 1.1:** Conventional and automated process of materials discovery experiments. 1) Production of samples, 2) evaluation of materials, 3) data analysis, 4) planning further experiments. While the conventional approach involves scientists for several tasks, the automated method excludes the humans-in-the-loop to enhance the throughput.

Recent advances in high-throughput synthesis experiments enable researchers to rapidly identify materials with desired properties and accelerate the materials discovery process [7]. Combinatorial synthesis and screening of large numbers of compounds allow for rapid exploration of compositional spaces [6], and statistical methods reveal patterns and correlations in the data that may not be apparent through traditional analysis methods. Additionally, the development of specialized hardware, such as automated deposition systems, enabled high-throughput synthesis experiments beyond combinatorial approaches and further accelerated material discovery experiments [7].

Accordingly, Figure 1.1 visualizes the process of conventional and automated material discovery experiments, split into the following essential steps: fabrication of sample(s), experimental material evaluation, data analysis, and the planning of further experiments. The conventional workflow integrates researchers for manufacturing, sample preparation, instrument operation, and subsequent interpretation of the data obtained from characterization techniques. Based on the data analysis results that identify prospective specimens, as conceptualized by the crosses, further target materials are manually defined for the next experimental series, and the process starts over. Thus, the pipeline's throughput is restricted by

several factors, such as the available workforce or merely the expertise to analyze the recorded data from the characterization techniques. As an alternative, a conceptual automated material discovery workflow increases the throughput of the experiments by eliminating the integration of humans in the pipeline. Samples are produced on a robotic platform and subsequently examined (1 & 2), while the data analysis (3) and planning of subsequent experimental series (4) are performed by mathematical models and computerized systems. However, the automated analysis of the experimental data remains a significant challenge [8, 9]. Thus, the conceptual, fully automated workflow has yet to be fully realized.

Therefore, it is essential to strategically select and employ characterization techniques that provide the most crucial information to reduce and simplify the data analysis effort. Although tension testing allows for determining key mechanical properties such as strength, flexibility, and fracture toughness, the preparation of the specific specimen is rather time-consuming and, hence, unsuitable for a high-throughput workflow. Alternatively, researchers can deploy non-destructive techniques that directly provide insights into atomic-scale properties from the produced samples. For example, diffraction techniques, such as X-ray or neutron diffraction, can offer information about a material's crystalline structure and packing of atoms, which infers other essential traits of the substance [10]. Efficiently packed structures generally exhibit enhanced mechanical properties due to their high density, while open-packed structures, with more space between atoms, might offer different properties, such as increased reactivity [5]. Thus, applying characterization techniques that enable the determination of properties on the atomic scale is crucial for automating the material discovery experiments.

While atoms and molecules, the building blocks of matter, are tiny, typically on the scale of picometers ($10^{-12}$m) and nanometers ($10^{-9}$m), their interaction with incident radiation or external vibrations enables the deduction of inherent attributes. Consequently, various characterization techniques have been developed to examine the properties of these microscopic structures. For example, X-ray diffraction (XRD) exploits the interaction of periodically arranged atoms with X-rays, which depicts several parameters of the underlying crystalline lattice and is a valuable tool for analyzing the crystal structures of materials. Consequently, XRD

instruments are commonly found in labs worldwide because they are versatile and widely applicable to many materials [10]. Similarly, techniques like Raman scattering and Nuclear Magnetic Resonance (NMR) spectroscopy are particularly capable of analyzing vibrational modes, shedding light on the dynamic behavior of atoms and molecules. Owing to their ability to determine key properties of the synthesized samples, these characterization techniques have become essential tools in the analysis of novel materials [11].

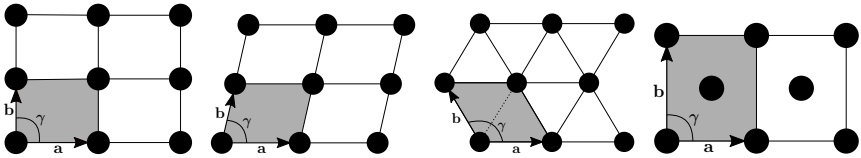
Nonetheless, diffraction and spectroscopy data analysis can be elaborate and time-consuming, requiring expertise in material science and data analysis. To address this challenge, researchers have increasingly employed machine learning techniques, primarily to diminish the dimensionality of spectral data, thereby improving its interpretability [12, 13, 14]. Furthermore, neural networks have emerged to identify crystalline phases from XRD patterns and Raman spectra, offering a fully automated evaluation of characterization data that eliminates the need for human intervention [15, 16, 17]. However, while the integration of neural networks has shown promise in accelerating the data evaluation procedure, it is essential to note that their application, so far, has predominantly been demonstrated in specialized tasks. The broader transferability of these automated approaches to diverse datasets, including materials that have not yet been discovered, remains a notable challenge in current research efforts. As such, there is a pressing need to explore and enhance the adaptability of these techniques for a more comprehensive and generalized application across a spectrum of scientific investigations.

Accordingly, this thesis addresses the inherent challenges of applying neural networks for data analysis in material discovery experiment pipelines. Thus, several methods are presented to facilitate model development, training, and application, thereby accelerating the data analysis procedure across various characterization techniques. Integrating these automated data analysis approaches represents a crucial first step toward achieving fully automated material discovery workflows, as illustrated in Figure 1.1.

To gain a comprehensive understanding of the novel methods introduced in this thesis, it is imperative first to grasp the foundational principles of crystalline substances and the characterization techniques utilized for acquiring the scattering and spectroscopy data. Consequently, Section 1.2 provides a thorough overview of crystalline materials, while Section 1.3 illustrates the interaction of samples with incoming radiation, which results in the measured characterization data. Subsequently, the manual approach for analyzing the measurements is demonstrated (see Section 1.4), and a synopsis of state-of-the-art methods for automated analysis of such techniques is provided in Section 1.5. Finally, the limitations of previous studies are stated (see Section 1.6), and the objectives of the thesis are summarized in Section 1.7.

## 1.2 Crystalline Structures

Elementary particles, such as protons, neutrons, and electrons, constitute the observable universe and all objects. Based on the elementary particles, various species of atoms are formed and are known as chemical elements, differing in the number of protons in their nuclei. Most of the substances found in nature are compounds of elements with ionic, covalent, or metallic bonds that keep the atoms of different species together [18]. In addition to the forces that result in strong bonds, atoms connect through weak forces, such as dipole-dipole interactions or the van der Waals force. The bonding forces dictate in which state of matter a material occurs naturally. If the particles are locked in place by strong bonds that allow only minor vibration movement of atoms, then a substance is denoted as solid. Although attracting forces are also prevalent in liquid substances, particles can move and slide past each other while remaining relatively close together. Therefore, liquids have a definite volume but not a definite shape, as they adopt the shape of the container. However, there is little free space between particles in liquid and solid substances, so they are not compressible. On the contrary, particles in gaseous matter interact weakly and occupy all of the available space [10].




**(a)** Rectangular or square Bravais lattice **(b)** Oblique Bravais lattice **(c)** Hexagonal Bravais lattice **(d)** Centered rectangular Bravais lattice

**Figure 1.2:** Types of Bravais lattices in two-dimensional space with their respective unit cells (highlighted) and the spanning vectors **a**, **b** plus the enclosed angle $\gamma$. [20]

State transitions occur when external properties are changed, e.g., temperature or pressure. For example, a solid substance changes into a liquid state when the supplied thermal energy amplifies minor movements of the particles to a certain extent that break the weakest bonds [19]. Similarly, gaseous and liquid substances can be cooled, so the movement of their particles becomes restricted until they are locked in place. When solids are formed, the particles are distributed in an equilibrium between attractive and repulsive forces. Most frequently, the optimal distribution of particles results in periodic arrangements: crystals. An example of this is halite, the naturally occurring crystalline state of sodium chloride (NaCl). Nonetheless, crystals can appear in all sorts of shapes and colors. Alternatively, when liquid materials are rapidly cooled so that particles do not have time to distribute to optimal states, atoms are randomly arranged, called an amorphous solid, which occurs most prominently in glass.

Crystals are periodic in one, two, or three dimensions and are theoretically infinite but, in practice, limited because of naturally occurring defects. The periodic structure is usually described as a lattice with elementary parallelepipeds, unit cells, that are repeated throughout the area or volume and are identical in shape and content [10]. Figure 1.2 shows the definition of different Bravais lattices with their spanning vectors **a** and **b**. Furthermore, the unit cell of each lattice is highlighted and described by the scalar quantities $a$ and $b$ plus the angle $\gamma$ that specifies the relative orientation of the vectors. By definition, each point within the Bravais lattice is representable through linear combinations of vectors with coefficients that belong to the set of whole numbers. Notably, these points in

the lattice are not limited to representing atoms or ions; they can also signify molecules (groups of atoms bonded together) [18].

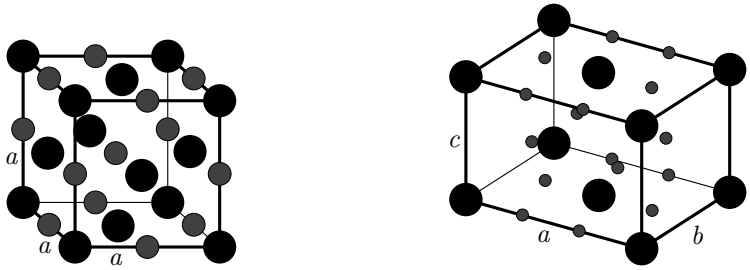
One way to discriminate between the different Bravais lattices is the relations and restrictions of the lattice parameters $a$, $b$, and $\gamma$ that specify the lattice type/system [20]. For example, the lattice vectors must be orthogonal for both rectangular and square unit cells ($\gamma = 90°$) but differ in the relationship of their magnitudes: $a = b$ for the square lattice (tetragonal system), $a \neq b$ for the rectangular (orthorhombic system). Similarly, the conditions $a = b$ and $\gamma = 120°$ apply for all hexagonal lattices, whereas the oblique unit cell (monoclinic system) has no restrictions. An additional discrimination is the position of the lattice points within the unit cell. For primitive lattices, particles (atoms, ions, or molecules) are located only at the corners of the unit cell, as illustrated in Figure 1.2a-c. Alternatively, Figure 1.2d shows a lattice with an additional point inside the unit cell, so the orthorhombic crystal system is split into primitive and centered rectangular lattices. Thus, a two-dimensional crystal can correspond to one of five distinct Bravais lattices and four unique lattice systems (monoclinic, orthorhombic, tetragonal, and hexagonal) [20].

However, solid matter has a volume, and thus, the periodicity in crystals usually spans three dimensions, which complicates the Bravais lattices and systems. In three-dimensional space, the lattice is described via three vectors, **a**, **b**, and **c**. Accordingly, there are six parameters to describe the lattice: the lengths of the unit cell edges $a$, $b$, $c$, plus three angles to describe the relative orientations, $\alpha$, $\beta$, and $\gamma$. Table 1.1 shows the seven lattice systems with their unit cell parameter restrictions and relationships for a three-dimensional space. While the two-dimensional monoclinic, orthorhombic, tetragonal, and hexagonal systems are complemented with two orthogonal angles to arrange their three-dimensional variants, there are additional systems (triclinic, trigonal, and cubic) with unique relationships and conditions. In addition to the primitive lattices for each system, three centered variants exist: body-centered with a point in the middle of the cell, face-centered with a point on every face, and base-centered with only two points in the middle of two parallel faces. Overall, there are 14 distinct Bravais lattices in three-dimensional space.

**Table 1.1:** Lattice systems and unit cell shapes in three dimensions [10]

| Lattice system | Unit cell shape/parameters |
|---|---|
| Triclinic | $a \neq b \neq c; \alpha \neq \beta \neq \gamma \neq 90°$ |
| Monoclinic | $a \neq b \neq c; \alpha = \gamma = 90°; \beta \neq 90°$ |
| Orthorhombic | $a \neq b \neq c; \alpha = \beta = \gamma = 90°$ |
| Tetragonal | $a = b \neq c; \alpha = \beta = \gamma = 90°$ |
| Cubic | $a = b = c; \alpha = \beta = \gamma = 90°$ |
| Hexagonal | $a = b \neq c; \alpha = \beta = 90°; \gamma = 120°$ |
| Trigonal/Rhombohedral | $a = b = c; \alpha = \beta = \gamma$ |

As previously noted, lattice points can represent individual atoms, ions, or even entire molecules, leading to complex crystal structures and the associated challenges in identifying their corresponding lattices. For instance, Figure 1.3a depicts the NaCl crystal with a face-centered cubic arrangement. The larger black dots in this structure represent sodium (Na) ions, which precisely align with the Bravais lattice points. Contrasting this, the smaller gray dots, representing chloride (Cl) ions, are positioned along the unit cell's edges and do not correspond to the Bravais lattice points. To streamline the crystallographic representation, the concept of a *basis*, which associates a specific group of atoms or ions with each lattice point, is often employed [18]. In the context of NaCl, the crystal structure becomes more discernible when the pair of sodium and chloride atoms is defined as the basis. Similarly, the Bravais lattice can be identified for the crystal structure of PuPt4, as visualized in Figure 1.3b. The larger black markers denote the plutonium atoms in this illustration, while the smaller gray dots indicate the platinum atoms. By designating the Pt-Pu-Pt atomic arrangement as the basis, the structure aligns distinctly with a base-centered orthorhombic lattice.




**(a)** Face-centered cubic crystal structure of NaCl [20]

**(b)** Base-centered orthorhombic crystal structure of PuPt4 [21]

**Figure 1.3:** Exemplary unit cell arrangements with differing content (black and grey points).

The crystal structures in Figure 1.3 exhibit another property that appears frequently in nature: symmetry. There exist several basic symmetry operations that can be described by their respective symmetry elements [10]:

- rotation by the rotation axis,
- inversion by the center of inversion,
- reflection by the mirror plane, and
- translation by the translation vector.

Accordingly, the NaCl structure in Figure 1.3a is defined by its definite symmetry operations. Primarily, it has three 4-fold rotation axes that pass perpendicular through the center of the cube's faces. The structure also showcases four 3-fold rotation axes, aligned diagonally between the cube's vertices, and six 2-fold rotation axes that pass through the centers of diagonally opposite, parallel edges. Furthermore, this face-centered cubic structure incorporates nine mirror planes and a center of inversion. By comparison, the orthorombic structure of PuPt4 also has the center of inversion. Still, a few mirror planes and rotation axes are missing due to the mismatch of unit cell edge lengths ($a \neq b \neq c$) and the difference between base-centered and face-centered lattices.

Additionally, basic operations can be combined and are then characterized by their complex symmetry elements: the (roto-)inversion axis for the mixture of rotation and inversion, the screw axis for rotation and subsequent translation, and the glide plane for the combination of reflection and translation [10]. By a combination of the 14 unique Bravais lattices and 32 distinct sets of crystallographic symmetry operations (point groups), the 230 essential space groups are formed (certain combinations results are not valid), which allows for the classification of all crystalline structures [18].

## 1.3 Characterization Techniques

### 1.3.1 General Diffraction of X-rays

Atoms and crystalline structures are too small to be observed using conventional light sources with wavelengths ranging from 400 to 700 nm, as their size is smaller by a factor of $10^3$ [10]. However, electromagnetic radiation with shorter wavelengths can interact with solid matter, as first demonstrated by Max von Laue in 1912 [22]. Most importantly, X-rays are used to analyze crystal lattices using effects such as scattering and interference. X-rays were first discovered by Wilhelm Conrad Röntgen [23] and have wavelengths between 0.1 and 100 Å, although only wavelengths in the range of 0.5 and 2.5 Å are commonly used in crystallography [10].

Most commonly in a laboratory setting, X-rays are produced using an X-ray tube, where high-energy electrons interact with a metal target. When the electrons collide with atoms of the metal target, they knock out orbital electrons from the inner electron shell of the atom. If electrons in the inner shell are missing, electrons from higher energy positions fill the vacancies and emit electromagnetic radiation when they change positions. The resulting wavelength of electromagnetic radiation depends on the target element used and the orbitals of the electrons. As a result, multiple characteristic wavelengths emerge from the X-ray tube, even for a

homogeneous target material, which is qualitatively presented in Figure 1.4. For electrons that transition from higher-energy orbitals, the emerging radiation has a higher energy, thus resulting in a lower wavelength. The three most prominent emission lines are K$\alpha_1$, K$\alpha_2$, and K$\beta$ with decreasing intensities in the same order. Typically, copper anodes are used in laboratories to produce X-rays, but molybdenum or cobalt targets are also used in specific settings [10].

When X-rays with a characteristic wavelength interact with matter, different effects occur depending on the properties of the impacted material. Firstly, X-rays are scattered by electrons of atoms, producing electromagnetic radiation spread in all directions. If the kinetic energy of the incident photons is not conserved, e.g., because of ionization, the wavelength of the scattered X-rays is increased, which is described as inelastic scattering. Alternatively, the energy and wavelength of the scattered X-rays match the incident beam for elastic scattering. Since X-rays interact with electrons, the scattered waves' intensity depends on the impacted material's electron density. Because not all radiation is scattered in the same direction, the intensity of the X-rays decreases with respect to the thickness of the material and might also be absorbed entirely [10].

When incident radiation interacts not only with individual atoms but also with atoms arranged in a periodic arrangement, the resulting scattered waves may share

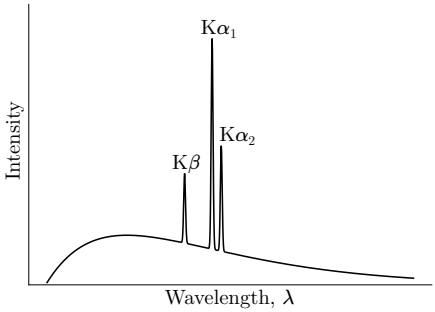



**Figure 1.4:** Qualitative emission spectrum of electromagnetic radiation from an X-ray tube. X-rays appear not only with a single wavelength but also with multiple characteristic wavelengths originating from different orbitals. Here, only the three most intensive emission lines are schematized and labeled. Based on [10].

a common orientation. However, owing to the spacing between the atoms, a path difference in the outgoing waves may arise. Figure 1.5 illustrates this for incident radiation under angle $\theta$ scattered on atoms arranged on parallel lines with distance $d_{hkl}$. The resulting path difference is $2 \cdot d_{hkl} \cdot \sin\theta$.

If the path difference aligns with the magnitude of the radiation wavelength, it results in either cancellation or amplification of the scattered waves. This phenomenon is described by the terms destructive and constructive interference. While visible light can not be used to record the diffraction pattern of a crystal, the wavelength of X-rays is comparable to the unit-cell spacings in crystals. Laue and his colleagues were the first researchers to come to this conclusion and reported the first-ever X-ray diffraction pattern in 1912 [22].

Expanding on this foundational work, Bragg's law [24] emerged as a crucial principle, describing the condition for constructive interference in the diffraction pattern:

$$n\lambda_{\text{X-ray}} = 2 \cdot d_{hkl} \cdot \sin\theta. \tag{1.1}$$

Consequently, for a given wavelength $\lambda_{\text{X-ray}}$ and a planar distance $d_{hkl}$, a diffraction peak of order $n$ is observed at angle $\theta$.

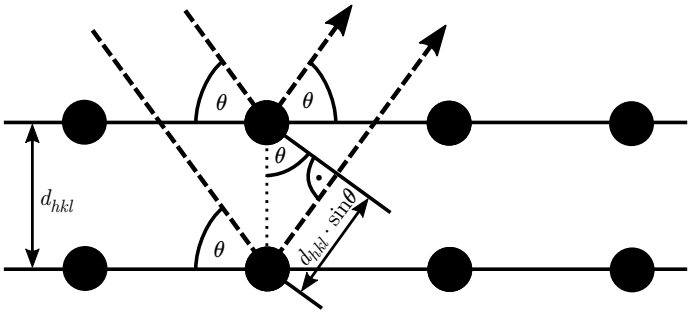
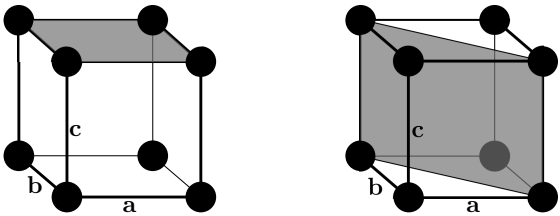



**Figure 1.5:** Diffraction of X-rays in a crystalline lattice with distance $d_{hkl}$ between the two illustrated parallel lines (planes in three-dimensional space) of lattice points. The incoming radiation occurs under angle $\theta$.

Families of parallel and equally spaced planes in a crystalline lattice are commonly described using Miller indices $h$, $k$, and $l$. They denote the reciprocals of where these planes intersect the crystal axes **a**, **b**, and **c**. As depicted in Figure 1.6a, the plane (001) intersects the c-axis at $c = 1$. Similarly, in Figure 1.6b, the illustrated plane (110) cuts through the axes at $a = b = 1$. As a result, the distance between neighboring planes is related to the lattice parameters [10]. In the trivial case shown in Figure 1.6a, $d_{001}$ is equal to the length of the lattice edge $c$. In general, the formula relating the interplanar distances with the lattice parameters is given below

$$\frac{1}{d_{hkl}} = \sqrt{\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2}}. \tag{1.2}$$

Consequently, a crystal has limited unique interplanar distances that satisfy the Bragg condition at specific scattering angles $\theta$. Notably, the intensity of diffraction spots from distinct planes remains unaffected by the presence of others so that multiple diffraction spots can be observed simultaneously in a single diffraction pattern (occurring at unique angles). The number of diffraction spots in a crystal pattern is intricately tied to its symmetry, as high-symmetry crystals exhibit fewer diffraction spots than their low-symmetry counterparts. The inherent symmetry in high-symmetry crystals leads to equivalent distances of Miller planes, so the resulting diffraction spots overlap. For example, in a cubic lattice where $a = b = c$, the distances of the planes perpendicular to the vectors spanning the unit cell are all equivalent $d_{100} = d_{010} = d_{001}$. In contrast, three distinct diffraction points




**(a)** Miller plane (001) **(b)** Miller plane (110)

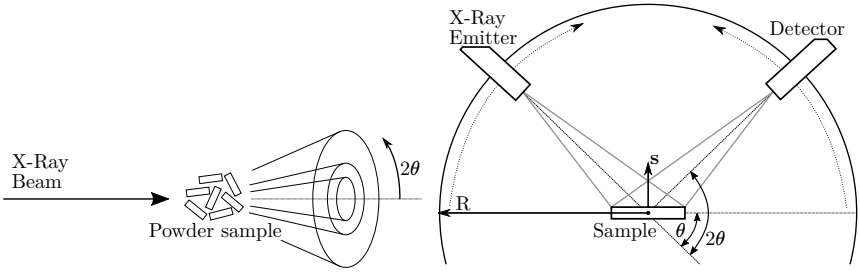
**Figure 1.6:** Two exemplary Miller planes in a monoclinic lattice.

can be observed for the $d_{100}$, $d_{010}$, and $d_{001}$ planes of an orthorhombic structure due to the diverging lengths of the unit cell edges [10].

Recording all diffraction spots, however, is an intricate task, as specific crystallographic planes must be precisely aligned to satisfy the Bragg condition for reflection. For example, planes oriented either perpendicular or parallel to the incoming X-rays do not yield a detectable reflection because there is no difference in the paths of rays scattered from these planes. Hence, in single-crystal diffraction, the sample is rotated during measurement to capture planes from all orientations; the resulting diffraction spots are registered on a planar detector [25]. Nonetheless, a comprehensive detector and sophisticated sample rotation mechanism are essential for this method. An alternative approach grinds the material into a powder, containing multiple crystals in random orientations, thus fulfilling Bragg conditions without necessitating sample rotation — this is elaborated upon in the subsequent section.

### 1.3.2 Powder Diffraction Patterns

Due to the challenges in preparing the specimen for single-crystal diffraction, powder diffraction is most commonly used to analyze crystalline samples [26]. In a powder specimen, multiple tiny crystals are randomly oriented and produce several diffraction cones that satisfy the Bragg condition. Figure 1.7a shows how a beam of incident X-rays produces multiple diffraction cones at angles $2\theta$ with respect to the originating orientation. Consequently, the Bragg-Brentano geometry enables a straightforward recording technique in which the X-ray source and detector are placed in a circular orbit $R$ and move in opposite directions to increase angle $2\theta$, as depicted in Figure 1.7b. The resulting intensity profile is a cross-section through all diffraction cones [10].




**(a)** Origin of diffraction cones with Debye-Scherrer rings for powder samples. From [26].

**(b)** Powder diffraction instrument employing the Bragg-Brentano geometry [27]. From [18].

**Figure 1.7:** Acquisition of X-ray powder diffraction patterns. The diffraction pattern appears as cones for samples that have been ground into a fine powder. Instead of recording the whole cones, scanning through the rings' cross-section reduces the required data volume while retaining the essential information.

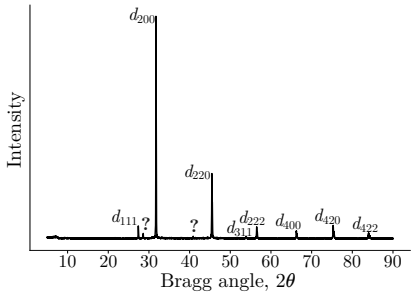



**Figure 1.8:** Powder X-ray diffraction scan of NaCl structure. From [28].

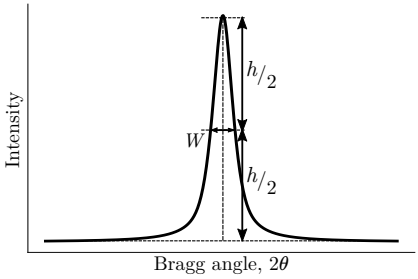



**Figure 1.9:** Characteristic appearance of peaks in powder diffraction patterns with properties height $h$ and width $W$ at half maximum.

The resulting one-dimensional powder diffraction pattern of a mineral sample is shown in Figure 1.8. In this example, the instrument recorded the diffracted intensities from 5 to 90 degrees $2\theta$ in steps of size $\Delta 2\theta = 0.01°$[1], so the measurement results in 8501 data points [28]. Thus, intensities are typically recorded as discrete counts, and, depending on measurement variables such as radiation dose

[1] The range of the scans, as well as the step width, can vary based on the specific use-case and configuration of the instrument.

and scanning duration per step, the absolute intensities vary considerably [10]. The recorded intensities depend on the respective atoms in the lattice that scatter the X-rays and are typically described by the structure factor, but other effects, such as temperature, can also affect the peak heights [18].

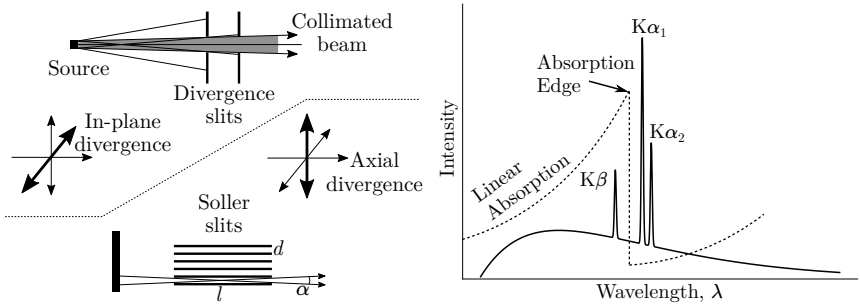
In the case of the one-dimensional signal of Figure 1.8, the diffraction pattern depicts the characteristic powder diffraction pattern of the NaCl structure from Figure 1.3a. For an XRD instrument with a copper radiation source, the sodium chloride sample yields nine characteristic diffraction peaks in the measurement range between 5 and 90° $2\theta$ that are attributed to unique distances $d_{hkl}$ in the crystal. Owing to differences in structure factors and plane geometries, the intensities of the recorded peaks vary, and specific peaks are barely discernible due to measurement noise introduced by the employed electrical components. Thus, the peak attributed to the $d_{331}$ planes, expected at approximately 74°, is not detectable in this scan, and the $d_{311}$ at 54.5° only marginally rises above the noise. Moreover, two ancillary peaks at 29° and 41° are present, as highlighted by the question marks, that do not result from the NaCl structure and are therefore attributed to impurities.

Furthermore, Figure 1.8 demonstrates another characteristic of XRD patterns: the diffraction peaks exhibit a broadened and modulated shape that diverges from the discrete interference effect formulation depicted by the Bragg equation. Though the Bragg equation pinpoints the ideal position of these peaks, corresponding to constructive interference of X-rays scattered by inter-atomic planes, the observed peak profiles in a pattern are often bell-shaped. The deviation from discrete peaks is partly attributed to the finite size of the crystals in the analyzed powder. In practice, imperfections and defects in the lattice are reasons why the Miller planes are not perfectly equidistant. Hence, the Bragg condition is satisfied for a small range of angles around the ideal distance $d_{hkl}$, and the diffraction intensities appear as broad peaks [10]. Moreover, the broadening is caused due to the interaction of the radiation with the optical components of the instruments. Accordingly, Figure 1.9 illustrates a typical peak shape with characteristic properties such as the peak height $h$ and the width $W$, which is measured at half the peak height and thus referred to as the full width at half maximum (FWHM).

The diffraction pattern of NaCl, as shown in Figure 1.8, is like a fingerprint for the mineral. Although other crystal structures have a matching arrangement of atoms in space and, therefore, identical space groups, the interatomic distances in the lattice depend on the atomic radii, so smaller or larger particles in the lattice shift the angles and positions of the diffraction peaks. Furthermore, different species of atoms feature varying scattering factors, so the intensities of the diffraction peaks differ, even for matching lattice parameters. Consequently, only the face-centered cubic crystal structure of NaCl generates a diffraction pattern with relative peak heights and positions, as shown in Figure 1.8. Yet, sodium chloride can adopt different arrangements due to variations in temperature, pressure, or other external conditions. Each distinct arrangement represents a different crystalline phase, which describes the precise atomic or molecular configuration for a given chemical composition. Therefore, the XRD pattern acts as a fingerprint for crystalline phases rather than chemical compositions [10].

As visualized in Figure 1.4, an X-ray tube produces radiation with multiple characteristic wavelengths in a laboratory setting. Hence, every Miller plane produces not one but three diffraction peaks if the electromagnetic radiation is not filtered. Similarly, focusing the beam on the sample is essential, or the diffraction peaks become broad and overlap. X-ray tubes, commonly used in laboratory settings, produce unfocused X-rays; therefore, powder diffraction instruments require additional apertures to condition the beam. Figure 1.10a shows two devices to restrict the dispersion of the beam in two directions: divergence slits for collimation of the beam and Soller slits with length $l$ and distance $d$ to limit axial divergence. In simple terms, the divergence slits focus the X-rays on the sample, while the Soller slits aim to improve the resolution of the diffraction pattern through parallelization of the waves (dispersing only with angle $\alpha_d$).

The characteristic emission profile of the X-ray tube can be filtered as shown in Figure 1.10b. Since electromagnetic waves with shorter wavelengths carry more energy, it is possible to filter higher frequency waves such as the K$\beta$ peak. As explained in the previous section, the absorption of X-rays occurs for some materials when high-energy radiation ionizes the material. For an X-ray tube with a copper anode, a Ni foil works best as a K$\beta$ filter that absorbs photons




**(a)** Focusing devices to correct axial and in-plane divergence

**(b)** Monochromatization of laboratory X-rays

**Figure 1.10:** Collimation and monochromatization of X-rays generated from an X-ray tube most commonly used in a laboratory setting. From [10]

with shorter wavelengths. At the same time, the energy of the K$\alpha$ radiation is insufficient to ionize the material, thus penetrating the foil. This effect is illustrated by the dotted line in Figure 1.10b, so both the K$\alpha_1$ and K$\alpha_2$ pass through the filter. In practice, the two K$\alpha$ wavelengths are very similar, and the peaks in the emission spectrum almost overlap, so filtering K$\alpha_2$ also reduces the intensity of the K$\alpha_1$ wavelengths. Hence, a monochromatization of pure K$\alpha_1$ is not feasible in most laboratory settings, and both K$\alpha$ shares are present in many available diffraction patterns [10]. In practice, the peaks generated from both wavelengths overlap for lower angles $2\theta$, and peak splitting is only observable for peaks that are located at higher angles, such as the peaks for $2\theta > 50°$ in Figure 1.8. All three devices, divergence slits, Soller slits, and K$\beta$ filters, are commonly placed in X-ray diffraction instruments. Both slits are typically located on both the emitter and detector site of the instrument depicted in Figure 1.7b [10].

Furthermore, the signal-to-noise ratio of the one-dimensional patterns is crucial to discriminate between low-intensity peaks and noise. Noise refers to recorded intensities that are not directly attributed to the analyzed sample but originate from other sources. Some of this unwanted intensity arises from X-rays reflecting off instrument components, such as the sample holder, rather than the powder sample itself. Additionally, the detector introduces uncertainties in intensity measurements due to inherent statistical effects. While longer measurement

durations can mitigate these statistical variations, an optimal experimental setup and specimen preparation are required to record XRD patterns with a sufficient signal-to-noise ratio.

### 1.3.3 Variation in Diffraction Patterns

While local imperfections and defects in the crystal cause broadening of the diffraction peaks, defects that affect the entire crystal introduce further pattern variations. Strained lattices emerge when the regular atomic arrangements of a crystal are disturbed or displaced, often due to external pressures, thermal effects, or the incorporation of foreign atoms. This strain-induced distortion results in peak shifts in the powder XRD signals, altering the characteristic patterns of phases. Similarly, there exist effects that change the scattering behavior of atoms and ions in the lattice, leading to height variations of the peaks. Consequently, the characteristic patterns of crystalline samples vary slightly, even for samples with identical phases.

Accordingly, Figure 1.11 shows the diffraction pattern for two samples of the mineral dolomite that differ in their peak positions and heights. Here, the pattern of sample B (orange) is offset on the y-axis to improve visibility. Apart from the presence and absence of some minor peaks, the most notable difference between the two patterns is that peaks at lower angles $2\theta$ are less shifted than peaks at higher angles. The shift of peak positions suggests the presence of a strained lattice and diverging lattice parameters between the two samples of the identical mineral. According to the Bragg equation (1.1), the change of lattice parameters shifts the peaks linear to $\sin(\theta)$. Hence, the position difference of corresponding peaks is more significant for higher angles $2\theta$.

The causes for divergences in diffraction patterns of samples from an identical phase are presented in Table 1.2. Specific properties of the crystal structure, the specimen, or the instrument's parameters affect either the peak positions, intensities, or shapes, which results in variations of the recorded interference pattern. For example, the defects and limited sizes in actual crystals cause the

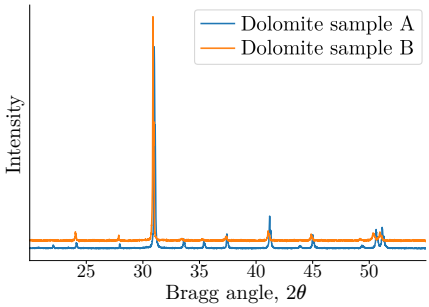



**Figure 1.11:** Two powder diffraction patterns for distinct specimens of the mineral Dolomite. To improve visibility, the pattern for sample B is shifted slightly on the y-axis. The patterns exhibit a modest mismatch in peak positions and intensities. From [28].

**Table 1.2:** Classification of properties, effects, and parameters that influence powder diffraction patterns, attributed to the crystal structure, specimen properties, and instrument parameters. Overall, the position and intensity of the peaks and the peak shape can be manipulated. Adapted from [10].

| | Crystal structure | Specimen property | Instrument parameters |
|---|---|---|---|
| Peak position | Unit cell parameters ($a$, $b$, $c$, $\alpha$, $\beta$,$\gamma$) | Absorption<br>Porosity | Radiation - wavelength<br>Instrument/sample alignment<br>Axial divergence of the beam |
| Peak intensity | Atomic parameters<br>Temperature factor | Preferred orientation<br>Absorption<br>Porosity | Geometry and configuration<br>Radiation - Lorentz polarization |
| Peak shape | Crystallinity<br>Disorder<br>Defects | Crystallite size<br>Strain<br>Stress | Radiation - spectral purity<br>Geometry<br>Beam conditioning |

broadened peak shapes, as observed in Figure 1.8 and conceptualized in Figure 1.9. The relation between the characteristic peak width $W$ and crystallite size $L$, is defined by the Scherrer equation [10] with the Scherrer constant $K$

$$L_{\text{crystallite}} = \frac{K\lambda_{\text{X-ray}}}{W\cos(\theta_{\text{peak}})}. \tag{1.3}$$

Thus, the Scherrer equation allows for deducing the underlying size of the crystallite grains in the sample, given the wavelength of the incoming radiation $\lambda_{\text{X-ray}}$, the position of the peak $\theta_{\text{peak}}$, and the measured width of the peak $W$. In practice, however, the FWHM is affected by the size of grains in the powder and other effects, as shown in the third row of Figure 1.2. Although the broadening of the peaks preserves the area under the curve [10], the analysis of diffraction patterns is complicated once neighboring peaks start overlapping. Thus, thorough specimen preparation and proper beam conditioning are necessary to produce informative signals.

Finally, the recorded diffraction intensities vary considerably if the diffraction experiment is not cautiously performed. Theoretically, the intensities are related to the scattering factors of the atoms in the lattice, so it does not only matter what atoms are present in the lattice but also their positions. Furthermore, the vibration of atoms in the lattice is accounted for by the temperature factor, as higher temperatures increase the vibrations of the particles in the lattice and, therefore, influence the scattering factors. However, temperature-dependent influences, as well as other effects such as the specimen's absorption or porosity and the radiation's polarization, cause minor variances compared to the preferred orientation effect that occurs for an ill-prepared specimen. As shown in Figure 1.7a, the powder sample produces diffraction cones, and Debye-Scherrer rings as long as all particles are randomly orientated in the sample. For a well-prepared sample, the orientation of the grains ensures that every $d_{hkl}$ for Miller-planes in the lattice is represented equally. However, if the crystallites do not represent an isotropic shape but instead are shaped like needles or plates, the grains are more likely to lie flat in the specimen, and therefore, some of the Miller planes are over- or underrepresented. Consequently, this effect has the potential to change the relative intensities of the diffraction pattern completely, so the formation of preferred orientations has to be prevented during the specimen preparation process.

### 1.3.4 Analogies to Spectroscopic Techniques

While XRD instruments record the diffracted intensity across a varying angular spectrum, denoted as $2\theta$, spectroscopic techniques, on the other hand, produce patterns of peaks and valleys across a scanned range of energies or frequencies. More specifically, the x-axis of the acquired data represents energy for X-ray or UV-visible spectroscopy, frequency in the case of Nuclear Magnetic Resonance or microwave spectroscopy, and wavenumber for techniques such as Infrared or Raman spectroscopy. Similarly to XRD, the specific positions and relative intensities of the peaks in these spectroscopic patterns serve as characteristic indicators for certain features of the materials or molecules.

Raman spectroscopy, for instance, serves as a tool for analyzing the molecules' vibrational or rotational modes, which are defined by the atomic bonds that constitute the substance. This technique contrasts the elastic scattering effect exploited in XRD, where incident X-rays excite electrons in a sample material, leading to the emission of X-rays with equivalent wavelengths as the electrons revert to their initial orbitals. Conversely, Raman scattering engages the inelastic scattering effect, which absorbs a portion of the incident energy. Consequently, the energy of the scattered radiation diverges from the incident waves, and the presence of specific wavelengths in the outgoing radiation corresponds to characteristics of bonds within the analyzed substance. Furthermore, Raman spectroscopy employs a laser to excite molecular vibrations, providing insight into the properties of molecules rather than inter-atomic distances. As a result, the wavelength used in Raman spectroscopy is typically three orders of magnitude greater than that of X-rays, such as the commonly used wavelength of 785 nm. Thus, XRD is primarily utilized for the analysis of long-range ordered structures, which are based on symmetries in crystalline materials, while Raman spectroscopy is more sensitive to short-range order and can be employed for the analysis of both crystalline and amorphous materials [29].

Figure 1.12 shows a Raman spectrum for the mineral dolomite on the right. Here, the signal shows several intensity peaks at various positions across the scanned variable, the Raman shift, that describe the vibrational and rotational modes of

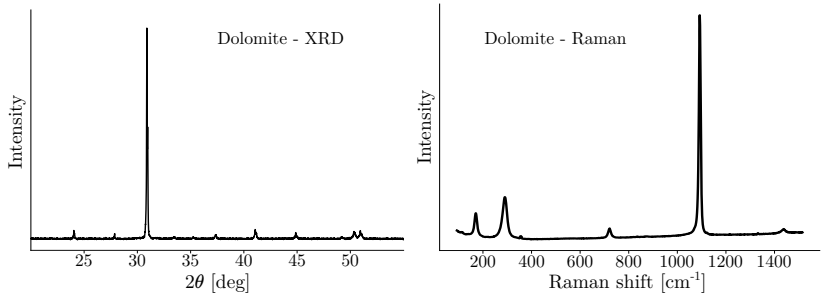
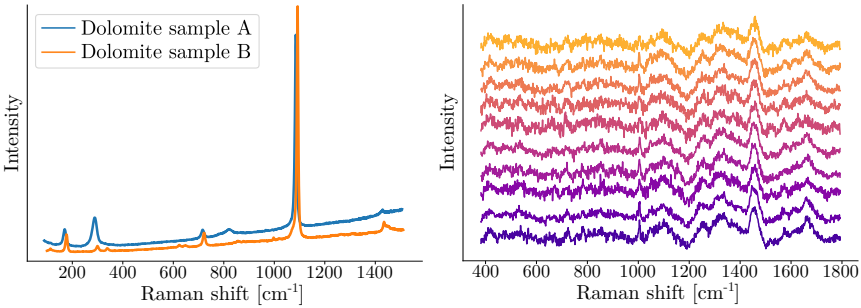



**Figure 1.12:** XRD pattern and Raman spectrum for the mineral dolomite, as acquired from the same specimen. Both techniques show intensity spikes that encode the relevant information concerning the peak positions and intensities. From [28].

the evaluated sample. For the identical sample, the XRD pattern is illustrated on the left in Figure 1.12, and both signals, XRD and Raman, exhibit notable similarities in their visual representation. Firstly, they share similar peak shapes due to the interaction of the measured radiation with optical components or other effects. Additionally, the essential information about the material or sample under study is encoded in the position and intensity of these peaks.

Another commonality in both XRD patterns and Raman spectra is the principle of superposition, which plays a crucial role when analyzing mixtures containing multiple substances. If a sample comprises several compounds, the observed pattern or spectrum is effectively a sum of the individual signals from each component. This means that the combined measurement can often be decomposed into its constituent parts. Nonetheless, the principle of superposition may not apply to all spectroscopy methods, especially if the presence of multiple components results in interactions that alter the measurement's outcome.

Another similarity of both characterization techniques is the shift of peak positions and intensities due to various effects that influence the sample. Mechanical strain, for example, affects the lengths and properties of bonds in crystals or molecules, which alters the recorded Raman shifts for the corresponding vibrational modes compared to an unstrained sample [30]. Accordingly, Figure 1.13




**(a)** Variation of Raman spectra for two dolomite samples. From [28].

**(b)** Variation of Raman spectra for identical species. From [31]

**Figure 1.13:** Measured Raman spectroscopy patterns that exhibit variation concerning peak position, intensities, and shapes.

shows the variation of different Raman spectra for various materials and substances. In Figure 1.13a, two measured Raman patterns for distinct samples of the mineral dolomite are shown, which differ slightly in peak positions, intensities, and shapes. Additionally, Figure 1.13b shows ten different Raman spectra for the bacteria isotope *Escherichia coli* that exhibit similar variation. Hence, while Raman spectra serve as fingerprints revealing the distinctive characteristics of the underlying sample's molecular bonds, associating these spectra with corresponding molecules can be challenging due to these subtle variations. Compared to XRD, Raman spectroscopy can also be used to evaluate organic samples that do not feature a crystal structure. The similarity in appearance and variation of the patterns from different characterization techniques results in the need for universal analysis methods.

## 1.4 Conventional Analysis of 1D Patterns

Building upon the theoretical background of signals acquired from techniques such as XRD or Raman scattering, the focus is now on the practical application of this knowledge for material characterization. Although the patterns for identical substances can vary due to artifacts, there is an underlying consistency

in the positions and heights of peaks in these patterns characteristic of certain substances. Therefore, this section delves into the methodologies employed for pattern analysis, primarily by utilizing reference data from various databases.

Figure 1.14 provides an overview of the conventional signal analysis process for identifying unknown samples. The measured signals contain artifacts that obscure the relevant information, so preprocessing methods are necessary to suppress noise or elevated baseline intensities in the signal. Because the characterization techniques have been established for centuries, many crystalline and organic samples have been examined and recorded to re-identify unknown specimens. Flawless reference patterns are stored in the database, or the material information is encoded and has to be transformed for comparison with the measured signals. Finally, the acquired information is compared with the database entries, and similarity metrics enable the matching with existing references.

Consequently, the following subsections explain the essential steps of conventional data analysis for one-dimensional patterns. First, Subsection 1.4.1 describes the preprocessing steps to obtain cleaner, more consistent signals ready for subsequent analysis. Then, Subsection 1.4.2 presents the conditioning steps to prepare the reference information from the databases for the matching. Finally, Subsection 1.4.3 defines metrics to assess the similarity of measured and reference patterns.

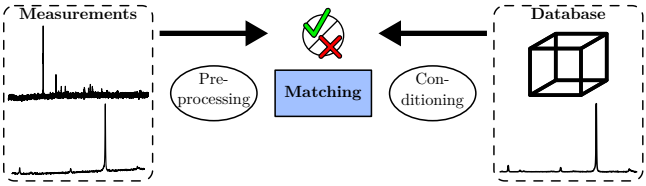
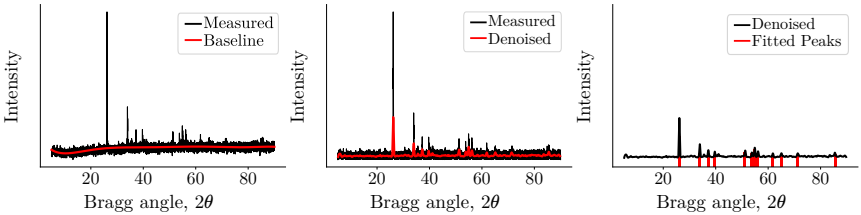



**Figure 1.14:** Process of identifying unknown specimens by matching the recorded patterns with database information using established methods. The acquired patterns are preprocessed to eliminate artifacts that hamper the matching procedure. Simultaneously, the database information is conditioned to match the properties of the recorded signals. Expert users can attribute the samples to database entries based on similarity metrics calculated in the matching step.

### 1.4.1 Preprocessing Methods




**(a)** Removal of the baseline intensity, which typically originates from substrate or fluorescence.

**(b)** Denoising of the scan to suppress intensity variations due to statistical effects when recording the data.

**(c)** Identified positions (as red bars) for the significant peaks in the signal.

**Figure 1.15:** Essential preprocessing steps to obtain the relevant peak properties (position, height) from the raw data, as shown on an exemplary XRD pattern of the mineral Manganite. Data from [28].

The detector in diffraction or spectroscopy analysis instruments records intensity values, but the acquired signal does not only contain the reflections related to the analyzed sample. For once, measured signals display fluctuations of the measured intensity values that have the potential to obscure the spectral or diffraction peaks, commonly known as noise. One of the most critical noise sources is detector noise, which arises from the electronic circuitry used to read out the detector, causing distortion in the recorded intensities. This introduces a statistical effect that causes the measured intensity values to scatter around the actual count. Additionally, radiation is scattered or reflected on the instrument's sample holder or other components if the beam or laser is not collimated correctly, resulting in elevated intensities across the measurement range [10].

One approach to compensate for noise in spectroscopic or diffraction data is to increase the acquisition time during the measurement. Increasing acquisition times can compensate for random and uncorrelated noise, as the longer measurement duration allows for averaging out of the fluctuations. Yet, as acquisition time increases, the sample may experience changes due to environmental factors, such as temperature or humidity, which can affect the quality and accuracy of the measurement. Additionally, longer acquisition times may not always be practical,

as they may be limited by the instrument's stability, the sample's availability, or the need for a quick analysis. Similarly, elevated baselines can be avoided through the use of elaborate components such as divergence or Soller slits. However, in a laboratory setting, the available instruments often limit pattern acquisition. Therefore, if available, careful consideration of the potential noise sources and the implementation of noise-reduction techniques are essential for achieving accurate and precise results in such sample characterization techniques [10].

To extract vital scientific information from measured signals that still contain artifacts, several preprocessing methods are crucial to prepare the noisy signals for the matching procedure. Accordingly, Figure 1.15 illustrates crucial preprocessing methods commonly applied when analyzing the recorded patterns on the example of an XRD signal for the mineral Manganite [28]. The raw diffraction scan is shown as the black line plot in Figure 1.15a, which exhibits increased intensity variations and an uneven baseline. However, the presence of complex, non-uniform background signals complicates the analysis process, as it poses a challenge in accurately quantifying the intensities of peaks. As a result, baseline correction is often necessary before analyzing XRD or Raman spectra to avoid misinterpretation of the results. Baseline correction can be performed through various methods, including polynomial fitting, spline interpolation, or Wavelet transforms [10, 32]. Polynomial fitting is a standard method that fits a polynomial curve to the data points in a selected region, subtracting it from the original signal. Alternatively, the Wavelet transforms method analyzes the signal at different frequencies, identifying baseline variations that can be subtracted.

Correspondingly, a fitted polynomial curve was used for the baseline removal step, as illustrated by the red line in Figure 1.15a. The corrected signal is then illustrated as the black line in Figure 1.15b, which still contains high noise levels. Similarly, various algorithms and approaches exist to suppress the noise in the signal while retaining the relevant information. One common approach to smoothing signals is to apply filters, which are mathematical functions that can reduce high-frequency noise while preserving the general shape and features of the signal. The choice of filter and its parameters depend on the specific characteristics of the signal and the type of noise present [32].

Figure 1.15b shows a denoised version of the diffraction signal in red. Here, a low-pass filter was used to eliminate the increased frequency noise in the signal while retaining the relevant diffraction peaks. The parameters of the filters have to be chosen manually, as no set of filter parameters exists that fits all the problems at hand [32]. Inappropriately chosen filters can also affect the shape of the relevant peak shapes, which complicates the following analysis process. The denoised signal still contains numerous peaks, some of which are leftovers from the suppressed noise. Hence, an essential step is identifying the relevant peaks characteristic of the analyzed specimen. Typically, peaks are identified by the first or second derivative methods, which identify peaks only in well-processed signals [10]. Alternatively, numerical peak search algorithms search for the highest intensity point and determine prominent points in the signal that satisfy defined parameters, such as the minimum intensity or the proximity to other peaks. Here, the identified peaks for the XRD scan of Manganite are displayed by the red bars in Figure 1.15c.

In addition to the described baseline correction, denoising, and peak search steps, more preprocessing steps exist, which are not necessary for the XRD pattern of the Manganite sample but can be relevant for other one-dimensional patterns. For example, due to the low-intensity nature of the Raman scattering process, the detectors in Raman scattering experiments are even sensitive to cosmic radiation, which produces narrow, additional peaks in the signals and must be removed. Similarly, non-monochrome X-rays result in duplicate reflections from families of planes in XRD patterns, so additional peak removal methods are necessary to obtain a clean diffraction signal. Information about additional preprocessing steps and more details regarding the presented data correction methods can be found in separate literature, including [10] and [32].

### 1.4.2 Conditioning of the Reference Data

While the previously described preprocessing steps address the preparation of the measured signals for the matching procedure, the reference information may also

require conditioning. If pristine scans match the measured signal with references, a prerequisite is that the length and step size of measured and reference spectra are consistent, so interpolation or cropping steps can be required. This is the standard procedure for the analysis of Raman spectra, for which several databases are available that contain Raman spectra of organic [33] or inorganic [34] materials and provide clean signals of the respective substances. In many cases, however, reference information is unavailable as pristine signals, as databases often store only the extracted data, such as the description of the identified crystal for XRD patterns.

Commonly used databases for the analysis of XRD data, including the Crystallography Open Database (COD) [35] and the Inorganic Crystal Structure Database (ICSD) [36], provide descriptions of distinct crystal structures derived from earlier research results. Determining crystal structures from XRD patterns involves a multi-step procedure, encompassing processes such as *ab initio* indexing and unit cell determination. However, the intricacies of these steps are not detailed within this thesis, which primarily focuses on the subsequent pattern matching, so a detailed explanation of this procedure can be found in separate literature, including [10]. To match this encoded information with the measured peaks, it is first necessary to compute the diffraction angles for the respective distances of inter-atomic planes in the crystal. Because the positions of the reflections in the diffraction patterns correspond with distances of parallel planes in the lattice, the Fourier transform, which allows for converting the spatial information of the crystal lattice into a frequency domain, is a method for rapid calculation of the exact positions [10]. To match the frequency of planes in the structure with the reflections in the diffraction pattern, it is furthermore necessary to consider the wavelength of the radiation in the XRD instrument to compute the exact positions of reflections using the Bragg equation (1.1) for all unique database entries.

### 1.4.3 Pattern Matching Procedure

Comparing the acquired signal with reference data from databases is commonly known as pattern matching. Typically, a single measured pattern exhibits similarities to multiple entries within the database. Consequently, methods are necessary to identify the most appropriate reference. Given that one-dimensional patterns may contain artifacts that cannot be eliminated through preprocessing steps, it becomes crucial to employ methods capable of accommodating differences between the measured and reference information. The selection of a matching approach depends on the specific characterization technique and the available reference information, leading to various methods with differing complexities. Notably, more sophisticated techniques often demand expert knowledge for effective application [10, 25].

This subsection briefly introduces, compares, and discusses the limitations of three primary methods for pattern matching: correlation coefficients, figure-of-merit, and full profile analysis.

#### Correlation Coefficients

The simplest methods for pattern matching involve comparing two one-dimensional vectors of identical length: the measured signal and a reference. Techniques such as correlation coefficients (e.g., Pearson $r_P$ and Spearman $r_S$) and error metrics (i.e., Mean Squared Error, MSE) fall into this category. The correlation coefficient quantifies the degree to which the two vectors vary, yielding a value between -1 and 1, and is commonly applied for analyzing Raman spectra [37]. A high absolute correlation coefficient value implies a substantial similarity between the two patterns. On the other hand, error metrics describe the total or average difference between corresponding points in the two patterns. Hence, a lower error score indicates a closer match between the patterns.

However, these techniques may not always yield reliable results when dealing with complex patterns. For example, measured patterns that are shifted on the x-axis

yield unfit error metrics when compared with non-shifted reference information, and only the ranked correlation coefficient is appropriate for evaluating data with such shifts [38]. Similarly, artifacts including broad peak shapes or duplicate reflections (from non-monochrome radiation) hamper the efficiency of those correlation and error metrics. In such scenarios, more sophisticated methods are required.

### Figure-of-Merit

As an alternative to basic pattern matching metrics, including correlation coefficients, the figure-of-merit (FoM) based approach offers a more nuanced approach by directly considering position or height variances of peaks. Unlike correlation coefficient-based matching, this approach compares discrete information rather than entire spectra. Hence, the FoM-based approach is applicable only when the reference information is provided in a discrete form. For instance, in XRD analysis, the peak positions and intensities can be derived from the database information, as explained in Section 1.4.2. Furthermore, it is necessary to determine the position and intensity of the peaks in the measured signal, as outlined in Section 1.4.1, which is not required for correlation coefficient-based matching [10].

The core idea behind FoM is to quantify the degree of alignment between the measured and reference peaks by calculating the differences in their positions or heights. With many peaks in the measured pattern, the initial crucial step is to establish correspondence between the peaks in the measured signal and their counterparts from the reference data. This correspondence can be established through strategies like nearest-neighbor matching or rank-based approaches. Once corresponding peaks are determined, differences in their positions or heights can be evaluated to determine the FoM value. In application, the FoM is typically scaled to 1 (perfect match), and higher FoM values indicate a better match between measured and reference peak information [25].

However, challenges may arise when matching the measured peak data with reference information. As the measured signal represents a superposition of

the fingerprints from all underlying substances, it potentially contains peaks that lack corresponding matches in the reference data. Furthermore, when a peak is present in the reference but remains unmatched in the measured signal, it could be obscured by baseline intensity or noise. To address this issue, software programs that analyze such patterns often incorporate customizable versions of the FoM. For instance, QualX, an established tool for XRD pattern analysis, provides options to fine-tune penalty terms related to peak position deviations or unmatched peaks [39]. Furthermore, it is essential to highlight that there is no definite FoM formula because closed-source applications, such as HighScore [40] or DIFFRAC.EVA [41], employ their variant of the metric.

Nonetheless, employing the FoM metric is not straightforward, as determining suitable parameters often demands expertise and a nuanced understanding of the sample. Appendix Section A.1 describes the phase identification procedure for the halite XRD scan illustrated in Figure 1.8 using QualX. Even with careful application, this example illustrates that the FoM metric may suggest phases not genuinely present in the sample, emphasizing the need to interpret the results thoroughly.

### Full Profile Analysis

While the FoM method primarily focuses on assessing individual discrepancies in the data, full profile analysis, mostly recognized as Rietveld refinement, takes a more holistic approach. Rietveld refinement employs a model encompassing sample properties such as unit cell dimensions, site occupancies, and instrument and specimen-dependent factors to reproduce the measured diffraction pattern. This model allows for simulating a comprehensive diffraction pattern, including accurate diffraction peak positions and intensities, realistic peak profiles, and the background intensity. The underlying model is then fine-tuned through iterative adjustments to align simulated and measured intensity values while considering constraints by the crystal symmetry of the underlying structure. Hence, if the lattice dimensions of the analyzed sample deviate from the corresponding reference in the database, causing mismatches in peak positions and diminishing the FoM

metric, Rietveld refinement addresses this mismatch by systematically adjusting the unit cell dimensions. This nuanced process holds the potential for a more precise alignment with measured intensities.

While Rietveld refinement is typically used to determine the underlying crystal structure, it can also be used to quantify the similarity of the measured signal with a reference. The consensus between the simulated and observed pattern is described by the Goodness of Fit (GoF) with indicators such as the Chi-Squared ($\chi^2$) or the Weighted Pattern Residual (Rwp) metrics [10]. In practice, full profile analysis has proven to be particularly effective for analyzing complex mixtures and non-ideal samples, where the overlapping of several reflections often complicates the interpretation of individual peak positions. Appendix Section A.2 presents the full profile analysis of an exemplary XRD scan in detail.

Despite its effectiveness, the refinement process comes with a high computational cost. Fitting sample parameters for thousands of reference materials demands substantial resources and is significantly more time-consuming than methods that involve metrics such as the FoM or correlation coefficients. The Rietveld refinement is inherently iterative, where the sequence of operations is pivotal, demanding a deep expertise in its application and a thorough understanding of the specimen in question. Therefore, the full profile analysis method introduces a considerable degree of complexity for adequate selection of parameters for refinement and restricting the range of valid values, posing a high risk of sample misclassification due to poorly selected parameter sets.

Accordingly, Table 1.3 presents an overview of the described pattern-matching techniques. The unique fingerprint within an acquired signal is compared to a list of candidates using matching metrics to identify molecules or materials. Typically, the candidate with the best match is selected as the corresponding substance for the measured pattern. Nonetheless, the described limitations underscore the frequent need for manual fine-tuning of the matching methods and thorough interpretation of the matching result to ensure accurate identification. Hence, an automated system that streamlines the matching process by computing the matching metrics

**Table 1.3:** Comparison of pattern-matching approaches for one-dimensional signals.

| Matching approach | Correlation coefficient | Figure-of-Merit | Full profile analysis |
|---|---|---|---|
| Reference information | full signal | discrete positions & intensities | crystal structure |
| Applicable domains | universal | universal | XRD |
| Metrics | $r_S$, $r_P$, MSE | FoM | GoF: $\chi^2$, Rwp |
| Computational cost | low | medium | high |
| Required expertise | low | medium | high |
| Limitations | shifted peaks, experimental artifacts | occluded peaks, identifying corresponding peaks | complexity |

and selecting the candidate with the best matching metric would suffer the same limitations.

## 1.5 Neural Networks for Diffraction and Spectroscopy Data

Representing an alternative to conventional methods, neural networks, a subset of machine learning and specifically deep learning, are a powerful approach for analyzing the acquired data. These mathematical models consist of interconnected layers with artificial neurons, including input and output layers, with a set of weights and biases for each neuron. Depending on the task, the network can predict one or many outputs for any given input, and the output is generated through a complex interplay of the input data and the underlying weights and biases. In the domain of supervised learning, pairs of input and desired output are available to adjust and fine-tune the parameters of the network in a training process so that the predicted output aligns with the actual output. Appendix Section A.3 explains the functionality and modules of the neural networks in more detail.

Neural networks have successfully been applied to achieve state-of-the-art results for automated data analysis across various domains, including the field of image recognition. In this context, images serve as inputs and display diverse

objects in various sizes and orientations, ranging from animals and cars to plants. The network's output layer features a neuron for each unique object, generating probability estimates for the presence of these objects in a given image. To address variations in object appearances, the network learns to filter out irrelevant elements in the images, such as backgrounds and noise, while detecting robust features that allow for identifying each object. In image classification tasks, neural networks consistently demonstrate superior performance on image recognition benchmarks, outperforming alternative machine learning methods and, in some cases, even surpassing human capabilities.

Similar to the challenges that occur in the field of image recognition, one-dimensional spectra, and diffraction patterns present unique fingerprints that exhibit variability in appearance, often obscured by background elements and noise. Accordingly, Table 1.4 presents a list of studies that applied neural networks for the automated analysis of Raman and XRD scans, ordered by publication date. Accordingly, the first documented usage of neural networks for analyzing one-dimensional patterns was demonstrated in the work of Park et al. [42], in which a model was successfully trained to classify XRD patterns by their structural symmetries (i.e., space groups). Since then, numerous distinct neural network models have been developed, each tailored for specific tasks, such as the analysis of complex multi-phase compounds [16, 43] and the identification of bacterial pathogens in organic samples [31].

In developing neural networks for analyzing one-dimensional patterns, a significant challenge lies in capturing the inherent variation in the fingerprint of each substance within the training data. For instance, in XRD, Table 1.2 categorizes various effects impacting peak positions, intensities, and shapes, necessitating the inclusion of such variances in the training data. Consequently, many networks designed to analyze XRD scans are trained using synthetic diffraction patterns generated based on entries from crystallographic databases. As an alternative approach, particularly in domains where simulating patterns is not as straightforward as in XRD, an elaborate sample preparation procedure has been implemented to manually generate several specimens that capture the variations for each unique

**Table 1.4:** Overview of selected publications that apply neural networks to analyze diffraction or spectroscopy data.

| Publication | Domain | Description |
|---|---|---|
| Park et al. [42] | XRD | Classification of crystal system, extinction group, and space group (7, 101, and 230 unique classes, respectively)<br>Trained using simulated patterns from ICSD<br>94.99%, 83.83%, 81.14% accuracies |
| Liu et al. [17] | Raman | Distinction of 512 minerals.<br>Trained on measured spectra from RRUFF<br>93.3% accuracy |
| Oviedo et al. [44] | XRD | Classification of crystal dimensionalities (3) and selected space groups (7)<br>Trained using simulated patterns from ICSD<br>92.9% and 89.3% accuracies |
| Fan et al. [45] | Raman | Component identification in mixtures (167 classes)<br>Trained using generated mixtures from pure spectra<br>$\geq$ 98.8% accuracies |
| Vecsei et al. [46] | XRD | Distinction of space groups (230)<br>Trained using simulated patterns from ICSD<br>76% accuracy |
| Ho et al. [31] | Raman | Distinction of bacterial pathogens (30)<br>Acquired measured spectra for pathogens<br>82.2% accuracy |
| Wang et al. [15] | XRD | Classification of metallic organic framework phases (1012).<br>Trained using simulated patterns from Cambridge Crystallographic Data Centre (CCDC)<br>56.7% accuracy |
| Lee et al. [16] | XRD | Classification of multi-compound samples (38 unique phases).<br>Trained using simulated patterns from ICSD<br>98% accuracy |
| Szymanski et al. [43] | XRD | Iterative classification of multi-compound samples (140).<br>Trained using simulated patterns from ICSD<br>93.4% accuracy |

class. Hence, developing a neural network for analyzing one-dimensional patterns is only viable if training data representing the necessary variations can be simulated or a substantial number of phase-pure specimens can be produced to obtain measured scans.

Numerous studies have employed neural networks for automated data analysis, yet the reported metrics exhibit a notable variance. The primary performance metric, accuracy, quantifies the ratio of correctly identified samples to total samples. As illustrated in Table 1.4, the reported accuracies range widely from 54% to 99%. Notably, a distinctive aspect among these studies is the development of unique network architectures, individually evaluated with specific datasets corresponding to each application. This discrepancy raises questions about whether the reported architectures have been explicitly fine-tuned for the signals in each study. Moreover, it raises uncertainty about how these models can be effectively transferred to different measurements.

## 1.6 Open Questions

While neural networks have been effectively applied to analyze diffraction data and spectra, research in this area remains fragmented. Their application has mainly focused on re-identifying already known substances, leaving uncertainty about utilizing these networks to identify novel molecules or materials effectively. Accordingly, the following questions remain:

- In the realm of material discovery, particularly within high-throughput systems, the potential of neural networks to automate data analysis is a promising avenue. However, the challenge lies in identifying a practical and beneficial application for these networks. Given the diverse range of substances analyzed in such settings and the variability in system configurations, which employ distinct instruments for the characterization of the samples, a general concept is necessary to integrate networks into such systems.
- Neural networks require training data to adjust the underlying parameters and align predictions with the actual output. However, crystallographic databases do not represent novel materials, so it remains questionable how

these materials can be represented. Similarly, producing phase-pure specimens of those substances is not an option to acquire measured training data. Thus, there arises a need for a conceptual framework to represent novel materials in training data, allowing neural networks to be trained to analyze such signals.

- To ensure the training of robust neural networks, it is imperative to include the full range of potential variations within the training data. Consequently, an extensive set of exemplary patterns is required to train the network appropriately. This prompts the need for a strategy to generate a large-scale set of signals depicting novel substances and the diverse variations in their fingerprints for adequate training of a robust model.

- Prior research has introduced various neural network architectures tailored to specific datasets. However, the effectiveness of applying these networks to other measurements, especially those containing novel substances, remains uncertain. Consequently, identifying the optimal architecture for recognizing unknown substances within these patterns becomes imperative.

- In the realm of material discovery experiments, where diverse substances are analyzed under various experimental conditions and characterized using a range of techniques, substantial volumes of data are generated. Recognizing the inherent diversity of this data, utilizing multiple networks becomes necessary to accommodate the diverse nature of substances and signals encountered in these experiments. Hence, developing an end-to-end approach spanning data generation to network training is imperative for streamlining the application of neural networks to distinct diffraction and spectroscopy patterns.

- The performance metric of the developed neural networks varies considerably across the different studies. Thus, an assessment is necessary to determine how effectively the networks perform compared to manual analysis of acquired signals. This evaluation includes the accuracy metric and factors such as the required time and complexity of the approach.

## 1.7 Objectives and Thesis Outline

Based on the previously described questions, the central objectives of this thesis are:

1. The design of a novel concept for facilitating the application of neural networks in high-throughput material discovery platforms. While neural networks are undeniably valuable tools for accelerating data analysis or enabling fully automated exploration, their effectiveness depends on selecting an appropriate architecture and implementing a robust training procedure. The proposed novel concept addresses these challenges, presenting a solution that streamlines the utilization of neural networks for automated data analysis of diffraction and spectroscopic data from material discovery experiments.

2. The development of a concept for representing novel materials from material discovery experiments. Despite the inherent uncertainty in the outcomes of these experiments, the properties of the materials and molecules involved remain known. Consequently, a thorough analysis of experiment outcomes concerning input substances' properties and composition is imperative to represent the resulting materials effectively. These representations enable the generation of training data essential for instructing neural networks to identify these substances within complex diffraction or spectroscopy datasets.

3. The design of a conceptual framework to facilitate the generation of training data in large volumes. As measured patterns contain a diverse range of naturally occurring artifacts, the crucial task is to guarantee that the training data thoroughly represents the full range of variation that influences the unique fingerprints. Consequently, the formulated framework enables the simulation of varied patterns for novel substances. It serves as a critical resource for training neural networks, equipping them to analyze and interpret real-world measured data effectively.

4. The development of a unified neural network architecture capable of analyzing various material characterization signals generated from experiments. Compared to established network structures that are tuned for specific datasets, this architecture should be optimized for signal analysis in general and exceed the performance metrics of the models presented in previous studies. Furthermore, the novel architecture should combine accurate predictions with minimal computational demands, facilitating seamless integration into existing systems.

5. The practical implementation of the innovative concepts and frameworks. To ensure accessibility and adaptability across diverse applications, the novel methods and associated tools are intended to be made available in open-source repositories. This initiative aims to foster collaboration, transparency, and the widespread adoption of the developed approaches within the scientific community.

6. As a practical demonstration, the developed concept is applied to automate the analysis of XRD data from distinct material discovery experiments focused on enhancing battery materials. The automated analysis approach is compared to manual data analysis, considering factors such as accuracy, complexity, and processing time, providing a comprehensive evaluation of the proposed method's effectiveness.

Accordingly, the thesis is structured as follows: Chapter 2 presents a framework to employ neural networks for automated data analysis in high-throughput systems. The framework includes an innovative concept for generating the training data, explained in detail in Chapter 3. Chapter 4 summarizes the commonalities and differences of network structures presented in previous studies to analyze unique datasets and introduces a universal dataset to assess the fundamental differences of these architectures. Based on the benchmarking results, the principles for the efficient design of neural networks are formulated, ensuring optimal applicability across various tasks and datasets.

To illustrate the utility of the introduced concept, Chapter 6 demonstrates the application of the framework for the automated analysis of various XRD datasets,

which were generated in experiments targeting the development of new battery materials. Additionally, the development effort for such a network is evaluated, comparing its accuracy and time efficiency against manual data analysis. Finally, Chapter 7 summarizes the contributions and provides an outlook on potential further research subjects.

# 2 Novel Substance Identification Concept

## 2.1 Material Discovery Experiment Analysis

In material discovery experiments, the primary objective is to identify combinations of precursors and experimental parameters that yield substances with advantageous properties. For instance, steel is a material of longstanding prominence with various applications. Fundamentally, it is an alloy composed of iron and carbon. Research has revealed that integrating various elements, like manganese, nickel, chromium, and vanadium, into steel can considerably enhance its key properties, notably toughness, strength, and corrosion resistance. Consequently, conducting experiments is essential to determine the appropriate combination of elements and the experimental conditions leading to structurally beneficial structures [47].

Accordingly, combinatorial synthesis is acknowledged as an effective method for screening multi-compound chemical spaces [7, 48, 49]. This approach is characterized by the definition of a chemical space through a set of precursors, which collectively span the composition space. In various experimental series, unique combinations of these precursors are mixed, and the resultant samples are subsequently analyzed. Hence, this approach facilitates the rapid identification of compounds that may possess beneficial properties, thereby enhancing the efficiency of material discovery research.

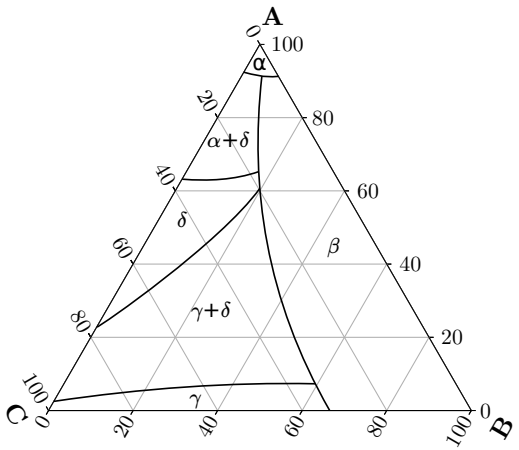



**Figure 2.1:** Exemplary ternary phase diagram with end members A, B, and C. Within this ternary composition space, the distinct phases $\alpha$, $\beta$, $\gamma$, and $\delta$ are formed, depending on the composition of the end members.

In experiments aimed at forming solid materials with crystalline structures, phase diagrams are crucial analytical tools in materials science, offering a visual representation of the stability of different material phases under various conditions of temperature, pressure, and composition [50, 51]. The diagrams provide clear insights into phase transitions, such as melting or evaporation, and are instrumental in determining the precise compositions for desired phases. Figure 2.1 presents an exemplary ternary phase diagram featuring arbitrary end members **A**, **B**, and **C**. This diagram displays the phase equilibria at constant pressure in a ternary system, and each corner represents a pure end member (100% concentration). The coordinates within the ternary diagram represent the specific proportions of each component in a mixture.

Researchers can discern intricate phase relationships within such a multi-component system in the detailed analysis of ternary phase diagrams. The diagrams typically delineate distinct regions where single phases are stable, such as the $\alpha$, $\beta$, $\gamma$, and $\delta$ phases in Figure 2.1. Moreover, they reveal areas where mixtures of phases coexist, for instance, regions indicating the simultaneous presence of $\alpha$ and $\delta$

phases. By interpreting these diagrams, scientists can understand the conditions under which specific phases or combinations are thermodynamically favored.

The exemplary phase diagram provides a clear representation of the solubility and phase behavior of components **A**, **B**, and **C** under specific conditions. It clarifies that at high concentrations of component **C**, only minimal amounts of **A** can be incorporated to maintain a phase-pure $\gamma$ structure. Exceeding these concentrations leads to the formation of an additional phase, $\delta$. In contrast, when component **C** is present in minor quantities, a wide range of mixtures of components **A** and **B** is permissible, consistently resulting in the formation of phase $\beta$. This detailed interpretation assists in understanding the complex interactions between the components, guiding the synthesis of materials with desired phase compositions.

Assuming phase $\beta$ is identified as having beneficial properties for a specific task, researchers can strategically utilize the information from the phase diagram to guide experimental designs. The diagram indicates the precise compositional ratios of components **A**, **B**, and **C** that lead to the formation of phase $\beta$. By targeting these specific ratios in their synthesis, scientists can optimize the likelihood of obtaining phase $\beta$ in their experiments. This approach significantly reduces trial-and-error experimentation, allowing for a more focused and efficient exploration of the chemical space.

Introducing an additional precursor, **D**, into this system complicates the phase diagram, as it extends beyond the ternary system to a quaternary one. This requires a more complex representation, often a three-dimensional diagram or a series of two-dimensional slices at different concentrations of **D**. Each slice or representation would show how the inclusion of **D** affects the stability and formation of the various phases, including $\beta$. Researchers can then analyze these diagrams to understand how **D** influences the phase behavior and identify the optimal combination of all four components to achieve the desired phase $\beta$. This expanded diagram provides a broader scope for experimentation, potentially leading to the discovery of more effective material compositions.

Phase diagrams for various combinations of components are readily available in scientific literature, providing a foundational understanding of material behavior in specific systems [21, 52]. However, further phase diagrams can be generated through combinatorial synthesis to explore uncharted compositions or refine existing diagrams. This approach involves systematically creating an array of specimens, for example, with different ratios of components **A**, **B**, **C**, and potentially **D**. The resulting specimens are then analyzed to determine their crystalline structures, with XRD analysis being a primary tool for this purpose.

An exemplary study on determining phase diagrams is the work conducted by Velasco et al. [53]. Their investigation focused on complex multi-compound oxides, exploring systems with five components: Ce, Pr, La, Sm, and Y. Depending on the specific composition of these elements, the resulting materials exhibited either a Fluorite or Bixbyite structure, or in some cases, a multi-phase mixture was formed. To accurately identify the crystal structures present in these specimens, the research team employed XRD and Raman spectroscopy, enabling the precise characterization of the materials and their corresponding structural properties. The primary objective of this study was to identify high-entropy materials that show promise for use in battery technology. Notably, unique compositions within these systems resulted in enhanced properties, highlighting the potential for improved applications in energy storage solutions.

In addition to the combinatorial synthesis and experimental evaluations, computational tools offer a complementary approach to material discovery. Notably, Density Functional Theory (DFT) [54, 55] and Graph Neural Networks (GNNs) [56, 57] have emerged as powerful methods in this domain. These tools take a specific composition with a defined structure as input and then predict fundamental properties, such as stability. However, a detailed explanation of these computational methods would be beyond the scope of this thesis. The outcomes of such simulations, including theoretical predictions of material properties, are typically cataloged in databases, including the Open Quantum Materials Database (OQMD) [58] or the Materials Project (MP) [59]. These databases serve as a repository for the results of these simulations, providing a valuable resource of theoretical materials for further research and validation. Integrating these

computational approaches with traditional experimental techniques represents a significant stride in the field of material science, enabling a more efficient and comprehensive exploration of material properties.

Despite the advancements in computational tools for identifying novel materials, the necessity for experimental validation remains [60]. One critical aspect is the determination of suitable precursor combinations and experimental conditions required to synthesize the material in question. In practice, the process can lead to the formation of either stable intermediate phases or volatile phases that may hamper the development of the intended target structure. Additionally, there is the possibility that the material, while matching the expected composition, may crystallize into a different structure than that predicted by computational models. These scenarios underscore the importance of experimental work and the necessity of methods to accelerate the data analysis process to identify the synthesized structures.

For instance, Szymanski conducted a study on the production of a highly fluorinated disordered rocksalt material, which has the potential to enhance the capacity of batteries [61]. However, there was a possibility of forming a volatile phase during the synthesis process. Therefore, several precursor combinations and experimental conditions have been evaluated to synthesize the target material. Ultimately, Szymanski et al. successfully determined an experimental pathway for synthesizing the desired material; however, the resulting sample still contained impurities.

Accordingly, characterization techniques, including X-ray diffraction, which enable the identification of structures based on their unique fingerprints, are crucial in material discovery experiments. These methods play a key role in verifying the presence of desired structures within synthesized samples. Consequently, the use of neural networks for analyzing sample characterization signals emerges as a powerful tool to accelerate the data analysis process, which is currently a substantial bottleneck in material discovery workflows. However, training neural networks to analyze such patterns and spectra necessitates a comprehensive representation of all possible outcomes in material discovery experiments. A

substantial challenge arises from the fact that novel materials frequently are not represented in crystallographic databases, as their occurrence and exact properties have not been confirmed through experimental validation before. Hence, the experimental pathways to synthesize these materials often remain theoretical rather than established. Additionally, identifying potential alternative phases in these experiments demands an in-depth understanding of material systems.

Furthermore, the development of an appropriate neural network architecture is critical. Such an architecture must be capable of accurately identifying materials based on their fingerprints while simultaneously disregarding irrelevant phases and experimental artifacts. This gap highlights the necessity of a concept for developing and applying neural networks to this data. Alongside the complexities associated with training data and network architecture, the concept must be versatile enough to accommodate a wide range of materials and characterization techniques. Additionally, it should be user-friendly and easily implementable, considering that researchers conducting these experiments typically lack specialized expertise in the domain of deep learning. Nonetheless, integrating neural networks with existing systems represents a pivotal step forward in the efficient and precise discovery of new materials, complementing combinatorial synthesis and computational methods.

## 2.2 Novel Substance Identification Framework

Figure 2.2 presents the developed framework designed for integrating neural networks to analyze material discovery data. At its core is the domain-specific neural network model, trained to identify crystal structures in a specific experimental series using a particular type of characterization data (e.g., XRD). The framework is structured into two main phases: training and application. During the training phase, which is separate from the sample production and data acquisition platform, the network undergoes a model optimization process to determine a set of model parameters $\Phi$ that achieve optimal model performance.

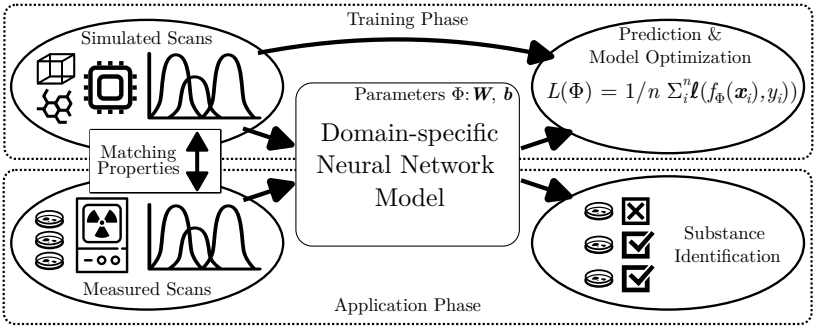



**Figure 2.2:** A unified framework to apply neural for uncovering novel substances in material discovery experiment data. To train the model, diffraction data or spectra are simulated to represent the target material. Using the generated training data, the parameters of a neural network $\Phi$ are optimized using a loss function $L$. Once a robust set of model parameters has been identified, the neural network can be applied to identify distinct substances in the measured scans.

Once the model is trained, the framework facilitates the automatic identification of specimens by their unique fingerprints. Due to the requirement of developing a distinct network for each characterization technique and target structure, the framework emphasizes efficient network design and training, ensuring it does not become a bottleneck in the data analysis process. Consequently, this framework is designed for seamless integration into existing high-throughput material discovery platforms and is versatile enough to handle diverse datasets.

The central aspects of the presented framework can be itemized as follows:

- **Simulated scans as training data**:
  A limitation of applying neural networks to measured diffraction and spectroscopy data is the availability of analyzed scans to train the models. Therefore, simulated scans, matching the characterization technique domain and the scanning range of the measured signals, are employed as training data. By simulating scans, novel substances can be represented as long as their properties can be described in a way that allows for the simulation of corresponding fingerprints. Moreover, the concept incorporates a generalized method for representing alternative phases within the

simulated signals, eliminating the need to outline every potential outcome of the experiments explicitly. As a result, the neural network excels at identifying substances in spectra or diffraction data, even handling cases where the experimental outcomes were not as anticipated.

- **Unified neural network architecture**:
  As the central component for analyzing both simulated and measured scans, an appropriate network architecture is required to identify the samples correctly. Given its intended application across different data characterization techniques and for identifying a range of structures, the framework incorporates a unified network architecture. This unified approach eliminates the need to develop different types of networks for each specific task. Additionally, the network is intentionally designed to be lightweight, eliminating network modules commonly found in established network architectures that increase the computational complexity without improving the performance proportionally. This streamlined design removes the need for specialized hardware to train the networks, facilitating integration into existing material discovery systems without requiring significant modifications.

- **Robust model training**:
  A robust training procedure is required to obtain a model that performs well on the measured data. During the training phase, the parameters of the network $\Phi$ are adjusted to minimize the loss $L$, which quantifies the discrepancy between prediction and label for the simulated samples. Therefore, guidelines are formulated to determine the appropriate loss function $\ell$ and metrics to assess the model's performance during training.

- **Application to measured scans**:
  After optimizing the network's parameters with simulated training data, the model can be applied to identify substances of interest in the measured scans derived from material discovery experiments. Minimal data processing routines are applied to ensure optimal performance. Importantly, this process is entirely automated and does not require human intervention, facilitating an efficient and accelerated data analysis pipeline.

In summary, the framework includes all essential steps for applying neural networks to identify those specimens containing the target structure based on their fingerprint in diffraction data or spectra. As the framework integrates a unified neural network architecture and robust model training approach, the concept does not require expertise in the field of deep learning, thus enabling researchers from other fields to develop neural network models for integration into their specific material discovery platforms. The following sections thoroughly explain the framework's central aspects and the detailed steps for applying the concept to specific data sets.

## 2.3 Framework Structure

Several prerequisites are essential to develop a domain-specific neural network model, which requires a thorough understanding of the samples and data involved. Firstly, an explicit knowledge of the domain of characterization data and measurement properties, such as the scanning range and step width, is necessary. Secondly, it is crucial to determine first which material needs to be identified. This could involve a description of the crystalline structure and the elemental composition so that corresponding diffraction patterns or spectra can be simulated. Furthermore, it is essential to know whether diverging experimental outcomes, for example, the presence of impurities, are crucial in the experimental series. Once the targets for identification and the types of scans are determined, the following steps can be undertaken to develop the neural network model for automated data analysis.

### 2.3.1 Simulated Scans as Training Data

The initial step in the framework is the generation of training signals for the neural network model. Prior studies have approached this by either simulating patterns based on entries from crystallographic databases [15, 16, 43] or acquiring pristine measurements of the substances targeted for identification [31, 45, 53]. Nonetheless, these methods might not be practical for material discovery experiments,

where various compounds can form, potentially diverging from existing reference materials or not being phase-pure. Therefore, an alternative approach is needed to generate the necessary training data, which does not demand extensive knowledge about the material system or the phases that may occur. This method offers a more accessible and adaptable solution for generating training data in complex material discovery contexts.

Accordingly, Chapter 3 describes the developed method for training data generation in detail. Central to this conceptual framework is the ability to identify materials based on their crystal structure, leading to a primary focus on the simulation of artificial XRD patterns in the framework. These patterns are efficiently computed for crystal descriptions through the Fourier transformation of the lattice [10]. Conversely, while Raman spectra are also helpful for material identification based on their unique fingerprint, simulating these spectra involves more complex *ab initio* calculations [62]. Consequently, tools for Raman spectra simulation are not as widely available or intuitive to use as those for diffraction patterns. However, recent research is focused on addressing this gap [63, 64, 65]. Thus, while the presented framework primarily focuses on XRD patterns, it can be extended to Raman spectra or similar characterization techniques, provided that tools are available to simulate the characteristic fingerprints accurately.

The conceptualized framework requires only a basic description of the target structure for generating training data. For structures identified through computational tools, this description can be obtained from databases cataloging theoretical materials, such as the OQMD [58] or the MP [59]. It is recognized that these described materials potentially exhibit variations in lattice dimensions when synthesized, and the data simulation approach is designed to accommodate these differences. In cases where experiments focus on material systems comprising multiple components, defining the target phase is adequate, such as the exemplary phase $\beta$ in Figure 2.1. The integration or substitution of different components typically leads to the expansion or contraction of the unit cell, impacting both the positions and intensities of the peaks in the characteristic fingerprints. However, the overall pattern typically remains similar. Accordingly, the framework is designed to represent these variances related to the target structure in the simulation

process, providing the training data to fit a neural network for identifying such diverse materials. As a result, these neural networks are equipped to recognize a wide array of novel materials, including those from combinatorial synthesis experiments or theoretical structures, enhancing their applicability in diverse material discovery scenarios.

Furthermore, the data simulation concept integrates an innovative method for representing alternative experimental outcomes without the need to explicitly describe the resulting phases or mixtures. In comparison to previous studies, this eliminates the need for a comprehensive understanding of the material system under investigation, thus streamlining the training data generation process. This innovative and exhaustive approach to data generation equips the neural network with a robust training dataset, enhancing its accuracy and reliability in identifying specific structures within the experimental variations.

### 2.3.2 Unified Neural Network Architecture

The framework features a unified neural network architecture, specifically tailored to accommodate the variations commonly found in diffraction patterns and spectra, and is effective across various types of characterization technique data. This is an improvement over the neural network structures developed in previous studies, which were developed for a specific dataset at hand. Therefore, Chapter 4 is dedicated to the development of the unified network architecture and the comparison with alternative network models. Consequently, two datasets are presented for the design and evaluation of an optimized neural network structure. These datasets unify consistent features of different characterization techniques and allow for a comprehensive evaluation of various network architectures. As a result, a model with an optimal set of parameters $\Phi$ is presented, which promises high performance in the analysis of measured signals from material discovery experiments.

### 2.3.3 Robust Model Training

Once the training data is generated and the exact network structure is determined, the network's parameters must be adapted to achieve optimal performance. To facilitate this, the generated set of signals is divided into a training set and a validation set. Typically, 80-90% of the examples are allocated to the training set, with the remaining samples comprising the validation set. This division is a standard practice in deep learning, where models are often over-parametrized and could otherwise memorize the training samples rather than learning robust classification rules applicable to unseen samples[66]. While the samples in the training set are used in the backpropagation procedure to fit the parameters of the model, the validation samples are not used to modify the parameters; instead, the model's performance on the validation samples is continuously monitored during training. Finally, training is halted once there is no further improvement in performance on these validation samples, ensuring the model's ability to generalize beyond the training data.

In this context, a *sample* is defined as a pair consisting of a signal (a one-dimensional vector describing intensities) and a label (a singular value specifying the corresponding class). The primary task of the network in this context is to identify fingerprints corresponding to the target substance, which corresponds to a binary classification task. Accordingly, the fingerprint is either correct (the target substance is present, label 1) or incorrect (the target substance is absent, label 0). As a result, the model is designed to predict a singular value for each signal, which is typically scaled between 0 and 1 using the sigmoid activation function.

To optimize the predictions of the model during the training process, the binary cross-entropy loss function is employed, as is common practice for binary classification tasks [66]. Given the known class affiliations of signals in the training set, the binary cross-entropy computes a measure of the quality of the alignment between predicted and actual labels for each sample. This value guides the backpropagation algorithm in fine-tuning the model parameters to enhance the network's prediction, as the goal of the training task is to minimize the loss.

For straightforward interpretation of the predictions, the output is commonly binarized [66]. This binarization process involves categorizing predictions greater than 0.5 as containing the target substance's fingerprint and those less than or equal to 0.5 as not containing it. These discrete predictions can then be directly compared with the discrete labels of the training data (either 0 or 1). Accordingly, the accuracy metric is calculated by assessing the proportion of instances where the model's predictions align with the actual labels, providing a clear measure of the model's performance, as introduced in Section 1.4.3.

While the principles and methodologies described here are well-established within the field of deep learning [66], their inclusion in this comprehensive framework is crucial. This ensures accessibility for researchers who may not be familiar with deep learning techniques and facilitates a broader range of scientists in effectively employing neural networks for their specific research needs. The outcome of the model training process is a finely tuned neural network characterized by a set of parameters $\Phi$ that yield optimal performance on the validation set. As the simulated training samples are specifically designed to replicate the artifacts present in measured scans, the fine-tuned model generally exhibits effective performance when applied to acquired signals.

### 2.3.4 Application to Measured Scans

In the final stage, the developed neural network can be applied to analyze measured data to identify substances in material discovery experiments. For the model to perform optimally, the instrument used for data acquisition must be correctly calibrated. While manual analysis has the capacity to adjust for minor calibration errors, the neural network's ability to recognize patterns is limited by the variation depicted in the training data. Accordingly, the trained model can consistently identify the fingerprint of the target material, as long as those deviations result in minor variations of the positions, intensities, and shapes of the peaks in the signal. Hence, larger deviations or systematic errors when capturing the signals could lead to unsatisfactory results. This highlights the importance of maintaining

precise calibration standards to ensure the accuracy and reliability of the neural network's identifications in practical applications.

Furthermore, it is essential for the length and step width of the measured scans to match those of the training data for optimal neural network performance. If a specific experimental series is acquired with a narrower scanning range than the training data, the missing intensity positions can be filled with zero values. Additionally, if the scans are acquired with a step size different from that of the training data, it is possible to adjust them through resampling or interpolation methods. However, if required, implementing appropriate data processing routines is mandatory to manage these deviations effectively.

Moreover, before inputting the measured signals into the network, it is crucial to scale them according to their minimum and maximum values. Because the training signals are generated without specific considerations for instrument configurations or data acquisition times, which can vary in actual experimental settings, the network is trained to process inputs scaled between 0 and 1. Consequently, this scaling process is crucial to normalize variations in absolute intensities that may result from different instruments or varying scan acquisition durations, thereby ensuring the network processes data consistently and accurately.

## 2.4 Framework Utilization

Several considerations are necessary to integrate the proposed framework into existing material discovery systems for analyzing various datasets. Due to the specialization of the trained neural network in identifying a specific type of substance and being tailored to a particular type of data from distinct characterization techniques, a unique model is required for each dataset. Accordingly, the following information is required to generate a new network model:

- **Characterization Technique and Radiation Source:**
  Different characterization techniques are employed to highlight distinct properties of substances, each yielding a unique fingerprint of the material

or molecule. Consequently, distinct models are required for interpreting the diverse characterization data. The choice of the radiation source is also crucial, e.g., what type of anode is used to generate the X-rays in diffraction experiments, as it significantly influences the resulting pattern.

- **Measurement Range and Resolution:**
  Determining the starting point, end point, and step width in measurements is another crucial aspect of training neural networks for anticipating the corresponding dimensionality of the inputs.

- **Target Material Description:**
  Formulating the unique properties of the target material is essential to train a network for identifying novel materials. Furthermore, whether the presence of impurities phases is acceptable has to be defined. Based on these properties, training data is simulated to train the models.

- **Simulation tool:**
  While the proposed framework is adaptable to various characterization techniques, a universal simulation tool for generating synthetic training data does not exist. Consequently, it is essential to identify or provide a simulation tool capable of producing the corresponding training data to effectively implement the automated data analysis concept with the given data.

Upon acquiring the necessary information outlined in the aforementioned points, the framework can be effectively employed. This framework encompasses a complete pipeline, starting from the generation of training data to the training of the model, culminating in a fully trained model that can be applied to the measured data without requiring manual intervention. Notably, the model is designed to be lightweight and quick to train, eliminating the need for extensive expertise in deep learning. As a result, this simplicity and efficiency facilitate the straightforward integration of the framework into various material discovery systems, enhancing their capabilities without demanding significant resource investment or specialized knowledge.

# 3 Generation of Training Data

## 3.1 Overview

The training of neural network models for application in spectroscopic and diffraction data analysis necessitates exemplary signals for fitting the parameters of the models. However, the characteristic fingerprint of novel materials is typically unknown, and experimental data depicting this fingerprint is not available, so generating examples using simulation tools is mandatory. Accordingly, the following Chapter unveils methodologies for accurately depicting material discovery experiment data in XRD patterns and spectra. Notably, the precise and efficient computation of such data holds a high degree of complexity, which could substantiate an independent thesis. Therefore, the presented framework utilizes existing simulation tools.

The Chapter is structured as follows: Section 3.2 analyzes established procedures for generating artificial XRD patterns or Raman spectra on a large scale to train neural networks to analyze such signals. Based on these methods, Section 3.3 highlights the limitations of existing approaches for applying material discovery data. Accordingly, a novel concept for generating an XRD training data set that overcomes these challenges is introduced in Section 3.4. The central component of the concept is the simulation of patterns that represent the target material; therefore, the concept is primarily designed to simulate XRD scans. However, Section 3.5 explains how this methodology can later be extended to the generation of Raman spectra and other spectroscopic techniques. Please note that parts of the current chapter are extensions to the work presented in [67].

## 3.2 Established Training Data Compilation Approaches

The goal of the automated data analysis approach is that the neural networks identify materials based on their fingerprint in diffraction data or spectra. Therefore, it is crucial that the training data contains all possible variations of the fingerprints, as demonstrated in various studies [44, 68]. Accordingly, several approaches have been utilized to compile effective training sets.

### Measured training data

An effective strategy for assembling a comprehensive training dataset involves acquiring multiple signals for each substance targeted for identification. For example, Ho et al. [31] prepared several samples of bacterial isotopes, subsequently obtaining their Raman spectra under diverse conditions. However, this approach is notably complex and time-intensive, mainly when the dataset must include a broad range of classes. Moreover, the instruments' characteristics can fundamentally influence the signal profiles, with each instrument adding a unique signature to the peaks. Therefore, relying solely on data generated from a single instrument raises concerns about the transferability and applicability of the dataset to spectra obtained from different instruments.

Alternatively, numerous studies [17, 45, 69, 70, 71, 72, 73] have utilized measured scans from the RRUFF database [28], which frequently offers multiple scans per material, to train neural networks for material identification. However, as outlined in Section 1.3.2, scans from the RRUFF database may include impurities, potentially training the networks to recognize incorrect fingerprints and resulting in misclassifications. Additionally, the representation of materials in the database is not consistent, leading to an imbalanced dataset. This disparity can prevent the effective convergence of the model, impacting its overall accuracy and reliability for materials that are less frequently represented.

### Augmenting scans

To address the challenge of limited measured training data, one method involves synthetically augmenting the existing data to enhance the variation within the training dataset. These augmentation techniques can include modifications like shifting the signals along the x-axis (in terms of $2\theta$ or wavenumber), adjusting the intensity of peaks (amplifying, diminishing, or eliminating them), or introducing noise and background intensities to the scans. Such augmentation methods are well-established for spectroscopy analysis techniques [74, 75], where materials are typically identified using exemplary scans and straightforward matching metrics, such as coefficient correlations (refer to Section 1.4). In the context of XRD patterns, a similar approach was adopted by Oviedo et al. [44] to expand a dataset for neural network training, demonstrating the adaptability of these techniques across different sample characterization techniques.

This approach to compiling a training dataset has been employed in various studies focused on the automated identification of materials via neural networks [44, 69, 71]. However, it necessitates the availability of a few exemplary scans for each material class intended for identification. These scans are typically obtained either through instrumental analysis or from databases offering exemplary spectra. Yet, this method proves inadequate for novel materials, where the appropriate set of precursors or experimental pathways remain unknown and are not represented in reference databases. As an alternative solution, Oviedo et al. [44] employed a data simulation technique, creating a set of artificial signals to which experimental artifacts were subsequently added, thus addressing the challenge of data scarcity for new materials.

### Training data simulation

A third approach in generating training datasets involves the simulation of training signals, which is particularly beneficial for XRD data. Simulating XRD diffraction patterns is straightforward, as Section 1.4.2 explains, making it an instrumental technique in this context. Accordingly, several tools are available for simulating

diffraction patterns [36, 76, 77]. In contrast, the simulation of training data is less established for Raman spectra and similar methods, mainly due to the need for complex *ab initio* simulations [62]. Central to this approach is the availability of material descriptions, typically provided in crystallographic databases [35, 36]. This method provides a viable alternative for generating robust training data, particularly in cases where experimental data are limited or unavailable.

Therefore, several studies presenting neural networks for analyzing XRD patterns utilized the training data simulation approach [15, 16, 42, 43, 44, 46, 68]. This approach can generally be separated into distinct steps:

1. identification of phases,
2. crystal structure variation,
3. peak intensity variation,
4. peak shape representation,
5. mixing of patterns,
6. addition of noise and background intensities.

In the following, each step is explained in more detail.

An appropriate set of phases must be selected from the available crystallographic databases, which commonly provide hundreds of thousands of references, albeit with varying degrees of data quality. Consequently, the first step involves filtering out unsuitable references for the intended purpose [42, 44, 46]. Additionally, since the training data often represents patterns from a specific material system, it becomes crucial to identify phases pertinent to this system and exclude irrelevant ones [16, 43, 68]. This process of selective filtration ensures that the resulting dataset is both relevant and of high quality, tailored to the specific requirements of the study.

The second step in the process involves the variation of crystal structures. As detailed in Section 1.3.3, the lattice parameters of a crystal structure are subject

to variation due to different factors, which show as position variations of peaks in the patterns. Lee et al. [16] addressed this issue by identifying all database entries corresponding to a specific material and simulating diffraction patterns for each individual entry. However, the frequency of database entries for each unique phase is inconsistent. While this method is effective for well-documented materials with multiple reports, it becomes impractical for specialized materials with limited representation in the database. Therefore, a universal strategy involves selecting a representative crystal structure for each material to be identified and then systematically altering the lattice parameters to artificially generate a range of patterns, accommodating the required variation [43, 68].

Another essential variation in the characteristic fingerprint of materials is the variation in peak intensities (see Table 1.2). Yet, only a few studies have incorporated this aspect. One effective method to depict this variation is considering the preferred orientation effect [43, 68]. By incorporating this effect into the training data, machine learning models can be trained to recognize variations in the position of peaks and their intensity variations. This ensures a more accurate and robust interpretation of the material's fingerprint by the models [68].

Furthermore, various factors contribute to the peak broadening effect in diffraction studies, leading to diffraction peaks displaying a more diffuse shape rather than sharp, distinct reflexes. Accordingly, the shape of powder diffraction peaks is commonly approximated using probability density functions to describe the statistical processes that cause the peak broadening. For instance, the centered Gaussian $G(x)$ and a Lorentzian $L(x)$ profiles with their characteristic parameters $\sigma$ and $\Gamma$ can be used to model peak shapes in diffraction patterns [78]:

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}, \tag{3.1}$$

$$L(x) = \frac{\Gamma}{\pi} \frac{1}{(x^2 + \Gamma^2)}. \tag{3.2}$$

Most important is the broadening related to the crystallite size, which is described by the Scherrer equation (1.3). The equation relates the crystallite size $L$ to the FWHM of the profile, which is connected to $\sigma$ and $\Gamma$ as follows:

$$\text{FWHM}_{\text{Gaussian}} = 2\sqrt{2\ln 2} \cdot \sigma, \tag{3.3}$$

$$\text{FWHM}_{\text{Lorentzian}} = 2\Gamma. \tag{3.4}$$

Building on this, the synthetic diffraction patterns are generated by utilizing these probability density functions to depict the peak shapes realistically. This method enables the modification of peak profiles to represent both changes in crystal size and distinct instrument configurations.

The previous steps have outlined the generation of characteristic patterns for unique phases. However, in practical applications, compounds typically comprise multiple phases. The principle of superposition applies to both diffraction patterns and spectra, as explained in Chapter 1, enabling a straightforward procedure for mixing patterns to represent such compounds. For instance, consider a compound composed of phases A and B. The generated patterns for these phases are denoted as $signal_A$ and $signal_b$, with each pattern's highest peak scaled to 1. A mixed scan, represented as $signal_{\text{mix}}$, can then be formulated as follows:

$$signal_{\text{mix}} = c_A \cdot signal_A + c_b \cdot signal_B \quad \text{with } c_A + c_B = 1, \tag{3.5}$$

where $c_A$ and $c_B$ are the mixing coefficients corresponding to phases A and B, respectively. This mathematical representation allows for creating composite patterns that accurately reflect the presence and proportion of multiple phases within a compound. Accordingly, several studies concerning the identification of materials in multi-compound mixtures have utilized this data generation approach [16, 43, 45, 68].

As the final step, noise and baseline intensities are added to the simulated patterns to account for the experimental artifacts. The Chebyshev polynomials $T_n(x)$,

which constitute a family of polynomials, are typically used as a representation for the background [16, 68] and are generally employed in signal processing to describe the effects of various kinds [79]. The polynomials of the first kind are recursively defined by the relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad \text{with } T_0 = 1, T_1 = x. \tag{3.6}$$

To generate diffuse backgrounds of various shapes, the Chebyshev polynomials up to the $N$-th order are summed using random coefficients $c_n$

$$P(x) = \sum_{n=0}^{N} c_n \cdot T_n(x) \tag{3.7}$$

and are evaluated for equally spaced points in the range [-1,1], matching the steps of the measured patterns.

Noise is typically simulated by drawing random values from the Gaussian $G(x)$ profile (Equation 3.1). To describe the level of noise in the signal, the signal-to-noise ratio (SNR) is a valuable metric

$$\text{SNR} = \frac{I_{\text{Signal}}}{I_{\text{Noise}}}. \tag{3.8}$$

Since arbitrary diffraction intensities are simulated, the maximum of the simulated intensities $I_{\text{Signal}}$ is typically set equal to one. Empirically, the typical SNR value lies between 10 and 35 for most XRD patterns [80], but low-count signals may require even lower ratios. Correspondingly, the mean $\mu$ and standard deviation $\sigma$ have to be adapted to match the level of noise intended to be depicted in the simulated patterns. Alternatively, Wang et al. [15] complimented their synthetic diffraction patterns with experimental signals, from which the peaks have been eliminated so only the baseline intensities and the noise remain in the signal.

The simulation approach for training data has considerably facilitated the application of neural networks to XRD data from diverse sources. Simulated XRD

scans have undergone thorough comparisons with their experimental counterparts, both visually and through systematic analysis, confirming that they are indistinguishable from each other [45, 46, 68]. Consequently, the performance of models trained on these simulated patterns translates effectively to measured data, yielding near-perfect accuracy metrics [16, 43, 45, 68].

## 3.3 Navigating the Challenges in Material Discovery Data

In material discovery experiments, various outcomes are possible, each with its own implications. The ideal scenario involves synthesizing the target structure without the presence of other phases. However, there are instances where the desired material may crystallize into an alternate structure, deviating from the anticipated experiment result. Additionally, the presence of other phases in the measured data is a common occurrence. Hence, the fingerprint of the target structure can be clearly detectable, overlap with patterns of other phases, or be entirely absent from the measured signal. Consequently, it is essential for the training data to depict all these potential scenarios.

As long as the key properties of the novel materials can be described, it is possible to simulate the fingerprint of the structures, so the training data simulation approach presents a straightforward method for tackling this challenge. These key properties specifically involve understanding the content within the corresponding unit cell and its arrangement, including consideration of the unit cell's dimensions. Alternative approaches, such as synthesizing pure specimens of the target material, acquiring their corresponding signals, and subsequently augmenting the training dataset, are not practical for novel materials. Although computational tools or phase diagrams often predict the existence of these novel structures, the methodologies for their successful synthesis are typically unknown prior to experimental attempts. Therefore, the training data simulation approach is without an alternative in material discovery data analysis using a neural network. While simulating XRD scans can be readily accomplished, tools for generating

characteristic Raman spectra of novel materials are currently unavailable. Therefore, this innovative approach should primarily focus on generating XRD patterns to facilitate the training of neural networks in identifying new materials. Importantly, the framework should be designed with adaptability in mind, allowing for the inclusion of alternative spectroscopic signals once tools capable of simulating accurate patterns become accessible.

Utilizing the training data simulation approach, Szymanski et al. [61] developed a neural network tailored explicitly for material discovery data. This model is designed to identify the desired structure based on its distinctive fingerprint in XRD scans. To accomplish this, they cataloged all possible phases that could be formed from the precursors used in their experiments and generated synthetic mixtures of these phases. Following the data simulation steps outlined in the previous section, they trained their model, which proficiently identified the target substance amidst various impurity phases. This highlights the practicality and effectiveness of neural networks for structure identification in material discovery experiments.

However, the approach adopted by Szymanski et al. [61] necessitates an initial comprehensive determination of all possible phases in their dataset, demanding an extensive understanding of the material system. This prerequisite, demanding prior knowledge of all possible phases, may not always be viable in material discovery experiments. In typical material discovery experiments, only the target structure and its corresponding fingerprint are known before performing the experiments. Identifying alternative forms of the target material or impurity phases generally necessitates conducting experiments and manually analyzing the acquired sample characterization data. Therefore, a more versatile and generalized approach becomes essential to accommodate the unpredictable and diverse nature of material discovery experiments.

Thus, it can be concluded that the data simulation approach, as outlined, is practical for generating training data. Still, it does not include a concept for accurately representing the range of possible outcomes in material discovery experiments. In general, the training data needs to depict both successful and

failed experimental results. This necessitates the development of a concept that can depict not only the desired structure but also impurity phases or alternative forms of the target material, including amorphous structures that manifest as diffuse diffraction patterns. Such a comprehensive approach ensures that the training data reflects the true complexity and variability of experimental outcomes, providing a robust foundation for the development of predictive models.

## 3.4 Novel XRD Pattern Simulation Concept

Figure 3.1 illustrates the developed framework for creating exemplary patterns, which are instrumental in training neural networks to identify novel substances within material discovery data. This concept constitutes a comprehensive, end-to-end framework for data generation that necessitates minimal inputs and autonomously produces realistic training patterns without the need for further human intervention. Central to the concept is the integration of existing simulation tools, such as pymatgen [76] or cctbx [77], for determining the exact positions and heights of the characteristic peaks. The first essential input to apply this novel framework is a detailed description of the target structure, typically available as a crystallography information file (CIF). Additionally, researchers must provide clear criteria defining what constitutes a successful synthesis outcome, particularly concerning the presence or absence of impurities. Subsequently, the framework generates patterns representing successful ("positive") and failed ("negative") experimental outcomes in synthetic XRD scans. These patterns are crucial for training neural networks to analyze and interpret such data effectively.

The primary objective of the overarching concept is to identify materials in measured data that have either been proposed by computational tools or novel material compositions situated in stable regions of a multi-dimensional phase diagram. Thus, the input to the data generation pipeline is either a computationally derived structure from a database or a structure that represents the stable phase targeted for production. However, several modifications to these foundational structures are required to reflect the natural variability encountered in experimental data

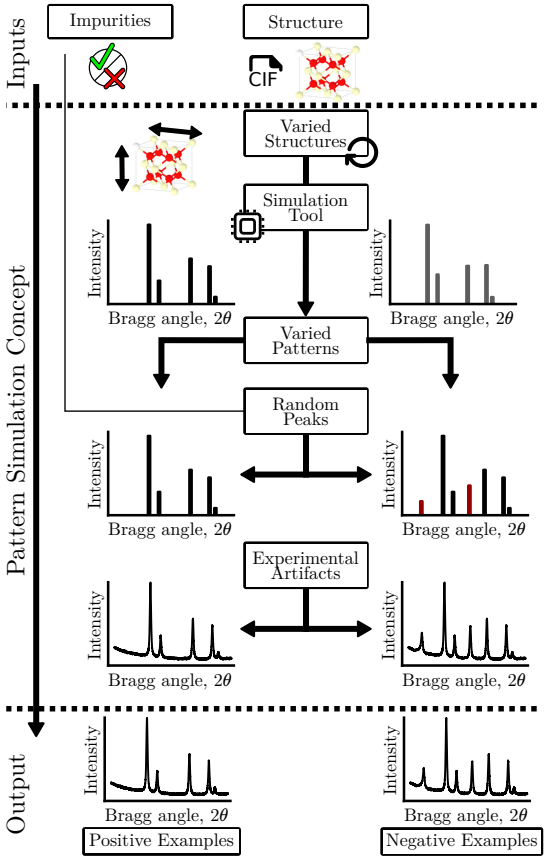



**Figure 3.1:** Framework for the generation of realistic powder diffraction patterns. Based on the provided material description (structure), exemplary signals are generated that either depict the correct fingerprint ("positive") or contain patterns that represent failed experimental outcomes ("negative"). The positive examples can either contain impurities or only depict phase-pure signals depending on the defined experimental conditions.

accurately. While computational tools are instrumental in confirming the stability of structures, they do not provide precise dimensions of lattice parameters. Similarly, the stable structure in a multi-dimensional phase diagram can display variations in unit cell dimensions due to incorporating species with different diameters, leading to either contraction or expansion of the lattice. Consequently, it is crucial to include these variations in lattice parameters within the training data to accurately capture the resulting changes in diffraction peak positions.

Furthermore, the diffraction peaks of the target structure can exhibit varying heights, especially for structures in multi-component material systems, where distinct species have differing scattering parameters. Hence, varied structures and corresponding patterns are generated based on a provided structure and a simulation tool. The generated patterns show variances in peak positions and heights while conserving the characteristic appearance of the fingerprint used to detect the target material in the measured data. Figure 3.1 provides a visual representation of these steps in the upper section of the figure. Accordingly, multiple *varied structures* are generated based on the initial description of the target's lattice. For these structures, the *simulation tool* is utilized to compute the accurate diffraction patterns, which are further augmented to reflect the intensity variations, resulting in multiple *varied patterns* that depict the fingerprint of the target material.

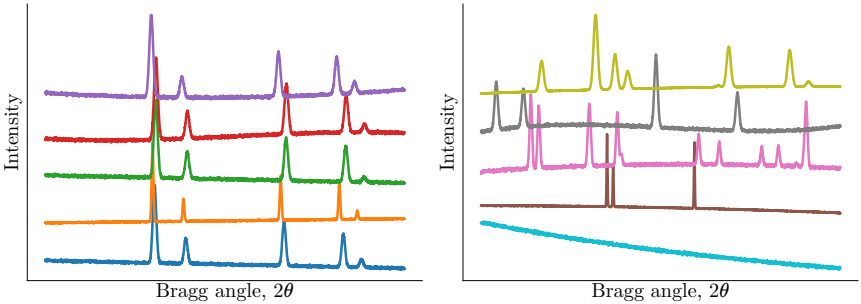
In the subsequent phase of the training data generation process, alternative materials are represented in the measured data using peaks with random positions and intensities. Previous studies have approached this by identifying the phases present in their data, simulating the characteristic patterns of these phases, and mixing these patterns with the primary structure's fingerprint for identification [61]. However, the methodology introduced here eliminates the need to explicitly determine these additional phases or simulate their precise patterns. As a result, this approach eliminates the requirement for specialized knowledge about the phases that may appear in the experimental data. Additionally, it saves time that would otherwise be needed for computing the accurate patterns using the simulation tool.

Because the training data has to depict both successful and failed outcomes of the experiments, the incorporation of *random peaks* plays a crucial role in depicting the varied potential results. Several peaks with random positions and intensities can be placed in an otherwise empty signal to generate unique fingerprints representing materials different from the target material. Similarly, the characteristic diffraction pattern of the target structure can be complemented with a small number of random peaks to simulate the presence of impurities in the sample. The addition of these impurities within the framework depends on additional input, which specifies whether impurities are considered acceptable outcomes of the experiments. Accordingly, impurities may be added either exclusively to the "negative" examples or to both the "negative" and "positive" patterns.

In the final step, *experimental artifacts* are incorporated into the generated signals. While this methodology has been well-established, its inclusion in the novel data generation framework remains crucial for generating realistic training data. The inclusion of experimental artifacts is achieved by convoluting peaks, which previously had discrete positions and intensities, with diverse Gaussian profiles. The profiles accurately replicate the various peak shapes observed in experimental data. A baseline intensity modeled by a Chebyshev polynomial is then added to the signals along with Gaussian noise.

Additionally, the framework accounts for experimental outcomes that result in the formation of amorphous materials. The absence of distinct peak shapes characterizes such materials. Therefore, samples containing amorphous materials are represented by including signals that consist exclusively of background intensities and noise without any superimposed peaks. This methodology is crucial for generating training data that enables the neural network model to interpret various experimental scenarios.

Accordingly, the developed concept for generating realistic XRD scans, as illustrated in Figure 3.1, produces a comprehensive dataset that contains both negative and positive examples for the fingerprint of the material to identify. Figure 3.2 displays a variety of signals generated using this novel framework. Here, the synthesized samples should contain the target material without impurities. Therefore,




**(a)** Generated patterns that represent positive experiment outcomes.

**(b)** Generated patterns that represent negative experiment outcomes.

**Figure 3.2:** Exemplary signals simulated using the presented framework for training data generation. Here, the presence of impurities has been specified as a failed experiment outcome.

the positive examples, as shown in Figure 3.2a, depict the characteristic fingerprint without additional peaks. Nonetheless, the simulated fingerprints exhibit varying peak positions, heights, shapes, diverse background intensities, and SNRs.

Conversely, Figure 3.2b illustrates signals corresponding to failed experimental outcomes. For example, the cyan signal depicts an exemplary pattern for a sample containing an amorphous material. The brown and grey patterns, distinctly different from the characteristic fingerprint of the target material, represent alternative structures that could potentially form during the experiments. Finally, the olive-green and pink signals show the target structure and complementary peaks representing impurity phases.

Therefore, the developed framework enables the rapid generation of XRD patterns representing various experimental outcomes. The framework incorporates the established data simulation steps outlined in Section 3.2 to generate realistic signals. However, the novelty of the framework can be summarized as follows:

- **Comprehensive approach**
  This framework is an all-including, end-to-end solution for generating training data. It is designed for efficiency, requiring minimal input and eliminating the need for manual intervention in creating exemplary patterns. It

incorporates all essential steps to simulate realistic XRD scans, proven to result in highly effective neural network models. This ease of use means that generating training data does not require extensive expertise.

- **Unified framework**
  The framework is adaptable to various material discovery experiments, whether the goal is to synthesize phase-pure specimens or samples with impurities. It introduces a unified concept that uses predefined criteria for successful synthesis outcomes as its input. This flexibility negates the need for additional modifications to suit specific experimental requirements.

- **Synthetic representation of phases**
  Unlike other approaches that necessitate identifying all phases within a dataset, this framework employs a more general method for representing alternative phases. Phases distinct from the target structure are simulated using arbitrary peaks with random positions and intensities. This versatility allows for the effective training of neural network models even in material systems where the possible phases in experimental data are not predetermined.

- **Inclusion of amorphous phases**
  Prior studies often overlooked the presence of amorphous structures, which typically yield patterns devoid of diffraction peaks. This framework addresses this gap by including negative examples featuring patterns that depict samples containing amorphous structures. This inclusion broadens the neural network's capability to recognize various experimental outcomes.

## 3.5 Extension to Spectroscopic Techniques

At the core of the developed framework lies the utilization of diffraction pattern simulation tools to accurately generate the characteristic fingerprint of novel materials. However, for Raman spectra and similar techniques, there exists a notable gap, as simulation tools for these methods are not readily available. In material

discovery experiments, Raman spectra analysis is particularly beneficial for the study of materials without a crystalline structure. Yet, tools that can be seamlessly integrated into the existing framework are not yet widely accessible, presenting an area for future development and enhancement.

Originally designed for XRD pattern analysis, the presented framework is versatile enough to be adapted for other characterization techniques, provided that suitable simulation tools are available. This flexibility allows for the generation of training data aimed at training neural networks to identify novel materials based on properties beyond the distances of parallel planes in the crystal, such as vibrational modes. A crucial aspect of creating an extensive dataset is capturing the variations in the target material's fingerprint, including changes in peak positions, intensities, and shapes. These key variations are integral to the framework, achieved through variation of the unit cell dimensions and randomly modifying the simulated intensities and probability distributions.

Introducing arbitrary peaks plays a crucial role in accurately representing structures that exhibit varying vibrational properties. The framework can be utilized to alter the unique fingerprint of a material by incorporating peaks at random positions, effectively simulating impurities in the sample, or generating patterns indicative of properties that differ from those of the target material. Consequently, realistic spectra can be generated with minimal effort, requiring simply the substitution of the simulation tool within the existing pipeline of the framework. This capability to adapt and simulate diverse spectra and patterns highlights the framework's comprehensive nature and its ability to support a wide array of characterization techniques.

# 4 Neural Network Model

## 4.1 Overview

At the core of the automated signal analysis framework, as introduced in Chapter 2, a neural network is employed for the categorization of the acquired signals. Numerous studies have demonstrated the application of neural networks to various XRD or spectra datasets, as summarized in Table 1.4. However, each of these studies introduced a unique network architecture that has been specifically tailored for analyzing signals in their respective datasets. Despite the unique nature of the signals investigated in these studies, certain similarities regarding the appearance of peaks in the spectra and patterns suggest that an optimized neural network architecture could achieve accurate predictions for all of these datasets. The versatility of such a unified neural network architecture would enhance the presented framework, enabling its seamless application across diverse datasets without necessitating the substitution or modification of networks based on dataset-specific characteristics.

Accordingly, the following chapter outlines the development and evaluation of a unified network structure. Section 4.2 describes the design of neural network structures in general and explains the configurable parameters. Based on characteristics derived from measured data, different configurations are examined, which results in the presentation of the optimized network structure that has been developed in the context of this thesis. Subsequently, Section 4.3 details an extensive comparison of the novel network structure with architectures presented in previous studies, highlighting the distinctive advantages and characteristics that distinguish the unified network. Finally, Section 4.4 explains how this versatile

network structure can be applied to diverse datasets generated in material discovery experiments, which does not require modifications to the network architecture. Please note that parts of the current chapter are extensions to the work presented in [81, 82, 83].

## 4.2 Neural Network Design

### 4.2.1 Parameter Consideration in Neural Network Design

Convolutional neural networks (CNNs) have demonstrated proficiency in analyzing noisy XRD patterns and Raman spectra [17, 42]. These networks have originally been developed in the domain of image recognition [84] and employ convolutional layers, in which kernels (filters) slide across the input to identify local, position-invariant features [66]. Rather than manually specifying these filters, the kernel weights are adjusted automatically during model training on the raw inputs. As a result, the convolutional layers are also useful for identifying the relevant features in diffraction patterns and spectra: peaks that are distributed over the entire length of the signal and obscured by noise and background intensities.

Figure 4.1 illustrates the conceptualized structure of a CNN. First, convolutional layers are employed for detecting the relevant features in the pattern. To capture the different kinds of features in the input, each convolutional layer employs multiple filters, resulting in the encoding of data into several channels. As a measure to downsample the inputs and identify features of larger sizes, pooling layers are integrated between the convolutional layers. Following the final convolutional and pooling layer, a flattening operation is introduced to reshape the encoded information, distributed across different channels, into a singular vector. Subsequently, one or multiple fully-connected layers are incorporated into the network, with the final layer providing the model's output. In contrast to the convolutional layers, designed to identify local patterns, the fully-connected layers are integrated to recognize features that are distributed across the whole size of the input [66].

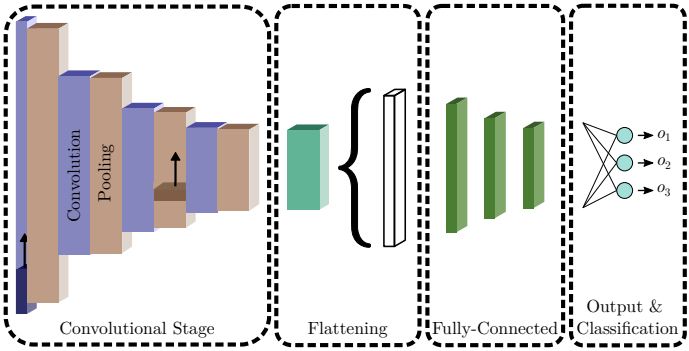



**Figure 4.1:** Schematic structure of a convolutional neural network.

Designing a neural network involves numerous possibilities, with each layer offering a set of hyper-parameters that can be adjusted. In convolutional layers, for example, the number of filters and the size of the kernels are parameters subject to adaptation. Additionally, CNNs can incorporate varying numbers of convolutional layers. The stacking of layers is another aspect that varies between CNNs: specialized architectures like VGG [85] or ResNet [84], for instance, stack multiple convolutional layers between the dimensionality-reducing operations (pooling). Furthermore, the Inception architecture [86] integrates several convolutional layers with varying kernel sizes, allowing simultaneous detection of features in different sizes. As a result, there is a multitude of configurations to consider just within the convolutional stage.

In light of the diverse configurations available, a straightforward approach could be the design of an over-parametrized network. For instance, an over-parametrized CNN might feature an abundance of convolutional layers, numerous filters per layer, and large kernel sizes. According to the universal approximation theorem, neural networks have the capability to learn any mappings between input and output given a sufficient number of parameters [87]. However, the inherent drawback of large networks lies in the demand for substantial resources during both training and application, coupled with the increased risk of overfitting. Thus, the challenge in determining an optimized neural network structure is to strike

a balance between equipping the network with enough parameters to detect the relevant features in the data while avoiding extensive parameter configurations.

Moreover, it is common to incorporate specialized modules into networks to counteract overfitting [66]. A notable example is dropout regularization, a technique that is typically integrated between fully connected layers. Interestingly, some studies in the field of spectral data identification have deviated from this standard configuration and instead integrated the dropout between convolutional layers [15, 46]. Another technique used to prevent the model from overfitting is batch normalization, which has also been used in networks presented for the analysis of Raman spectra [17, 31]. Hence, the task of designing neural networks extends beyond choosing suitable configurations for parameters and modules within the network; it also encompasses strategic decisions regarding the integration and placement of regularization techniques to address the issue of overfitting.

Accordingly, the following choices regarding the design of CNNs have to be considered:

1. the number of filters and size of the convolutional kernel in the convolutional layers,
2. the number of convolutional layers to identify complex features,
3. the exploration of the applicability of advanced image recognition concepts (VGG, ResNet, Inception) in the context of spectral data and pattern analysis (focusing on the strategic stacking of convolutional and pooling layers),
4. the integration and positioning of regularization techniques.

Therefore, a comprehensive assessment of various network configurations on an extensive dataset becomes imperative. One viable approach is to utilize the RRUFF database [28], which provides numerous measured XRD patterns and Raman spectra for diverse materials, for such an evaluation. However, the analysis of a halite XRD scan acquired from this database, as detailed in Section 1.4, exposes that some of these scans exhibit irregularities, as additional diffraction peaks are present in this pattern that are not attributable to the NaCl structure

reported by the database. Thus, an alternative large-scale dataset with reliable labels is necessary for the effective evaluation of the distinct configurations.

As an alternative strategy, a synthetic dataset tailored for the evaluation of network designs is introduced. This synthetic dataset is designed to mirror the characteristics observed in measured signals, drawing insights from a detailed analysis of the RRUFF database. Consequently, parameters identified through the synthetic dataset are expected to efficiently translate to the analysis of measured signals. Following the comprehensive assessment of various network configurations on this dataset, the thesis establishes the optimized network structure, subsequently integrating it into the novel data analysis framework.

### 4.2.2 Analysis of Measured Dataset Characteristics

To determine the characteristics of measured XRD patterns and Raman spectra, the corresponding data is downloaded from the RRUFF database and examined in detail. The database contains thousands of measurements contributed by different researchers using various instruments and measurement modalities. As a result, the database represents a wide range of variation, and the properties inferred from the RRUFF database are expected to be representative of measurements aimed at characterizing materials in a general context. The analysis is primarily aimed at determining key properties of the measurement data, such as scan length and peak width.

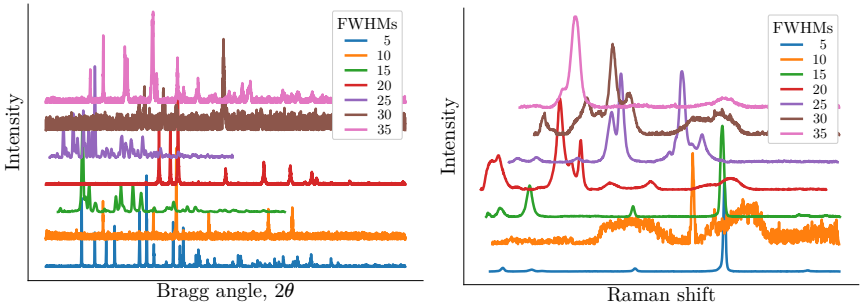
Examining the signals from the RRUFF database reveals a notable variability in the length of the scans. Considering only XRD patterns, the scans exhibit a range of 1000 to 9000 measurement steps. This variability is influenced partly by the specified measurement range tailored for particular materials. Moreover, the choice of different anode materials significantly impacts the XRD scan length, as shorter wavelengths cause diffraction peaks to appear at lower angles $2\theta$, which plays a role when specifying the measurement range. Additionally, the step width utilized in acquiring the scans dictates the number of measurement steps, with smaller step widths resulting in denser recordings of diffraction patterns, yielding

a higher number of data points. Similarly, the length of Raman spectra varies greatly with measurement ranges between 15 and 4000 $\mathrm{cm}^{-1}$, resulting in lengths between 600 and 8000 measurement points.

The measured intensities within those scans can exhibit substantial variations, spanning from hundreds to millions, attributable to differences in instruments and acquisition times. Furthermore, signals from both domains show multiple peaks, which are representative of the material under analysis. While certain signals showcase fewer than five peaks, others display substantially higher counts, such as XRD patterns of low-symmetry materials. Despite this variation in peak count, the shape and width of the peaks appear to be relatively consistent across all signals. Notably, the width of a peak in an XRD pattern is defined in terms of angle $2\theta$, and for Raman spectra, in terms of the Raman shift.

However, the width of peaks can also be expressed in discrete data points of the measured signal, facilitating a meaningful comparison between the two domains. For instance, if a peak exhibits a Full Width Half Maximum (FWHM) of $0.5^\circ$ in a diffraction pattern obtained with a $0.01^\circ \Delta 2\theta$ step width, this corresponds to an FWHM of 50 in terms of data points. This consideration of peak widths is particularly crucial in the context of inputting such data into neural network models. In the domain of analyzing spectral data, a vector detailing the measured intensities acts as input for the neural network, but it does not include information regarding the measurement steps. Thus, the deep learning models operate under the assumption of equidistant steps, even when the signal is acquired with a variable step width. As a result, the neural network may interpret the XRD peak with an FWHM of $0.5^\circ$ very differently depending on the step size of the data acquisition. If acquired with a step width of $0.01^\circ \Delta 2\theta$, the peak will be translated into 50 data points in width, while at $0.02^\circ \Delta 2\theta$, the same peak will be described in only 25 data points, even though it conveys the same underlying information.

Accordingly, Figure 4.2 displays a selection of XRD scans and Raman spectra from the RRUFF database. As explained earlier, the signals exhibit varying measurement ranges and lengths, highlighting the diversity of the dataset under investigation. Furthermore, the FWHM of the peaks has been determined for




**(a)** XRD scans from the RRUFF database.

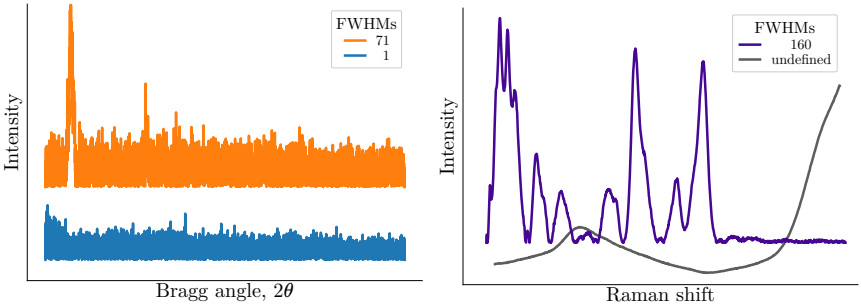
**(b)** Raman spectra from the RRUFF database.

**Figure 4.2:** Exemplary XRD scans and Raman spectra from the RRUFF database [28]. The database contains measurements from different sources, thus, the scans exhibit varying lengths, FWHMs, and SNRs. The colors indicate the characteristic FWHM of peaks in the signals.

each signal, as showcased by the differing colors. Thus, Figure 4.2 illustrates the overall consistency in the width of peaks observed in XRD patterns and Raman spectra from the RRUFF database. Notably, the peaks in the Raman spectra appear broader compared to their XRD counterparts, a characteristic attributed to the shorter length of the Raman spectra.

Nonetheless, it is crucial to thoroughly examine the scans available in the database to identify those with distinctly different properties. Figure 4.3 showcases several signals from the RRUFF database that stand out from the remaining scans based on their determined FWHMs. On the basis of this evaluation, XRD patterns with high and low FWHM values have been identified, which appear to be composed mainly of noise, as displayed in Figure 4.3a. However, as the signal does not allow for distinguishing relevant information required to identify a material from noise, such signals are effectively useless for the purpose of analyzing samples.

Similarly, there are Raman spectra in the database with outstanding FWHM values, as illustrated in Figure 4.3b. For some of these scans, no FWHM could be determined (gray), while others had particularly high FWHMs (purple). Upon closer inspection, the broad peak with the highest intensity in the purple spectrum appears to be composed of four separate peaks that overlap partially. Consequently, the actual FWHM value for each individual peak within this spectrum



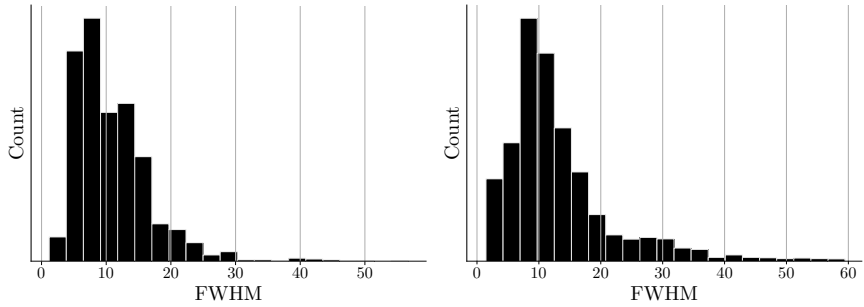

**(a)** XRD scans with exceptionally high or low FWHMs.

**(b)** Raman spectra identified through analysis of the FWHMs.

**Figure 4.3:** Signals from the RRUFF database [28] that have been identified based on their exceptional FWHM values. Such signals often consist entirely of noise or are the result of multiple peaks overlapping in the signal.

may be considerably smaller. Based on this investigation, it is crucial to recognize that the database includes scans with decisively divergent properties, including undefined peak widths or FWHMs that are exceptionally large or equal to 1. Hence, when determining the characteristics of scans within the database, such signals should be carefully identified and excluded from consideration.

Therefore, Figure 4.4 presents a histogram displaying the distribution of FWHM values in signals from the *filtered* RRUFF database. Thus, the frequency of different FWHM ranges is effectively represented by the histogram, which allows systematic determination of the peak width property of scans in the RRUFF database. As a result, the histogram shows that for XRD scans, FWHM values are predominantly centered around 10, with 99% of the scans having FWHM values smaller than 30. However, FWHMs of up to 60 can occur in measurement signals intended for analysis of various materials.




**(a)** Distribution of the FWHM values in XRD scans from the RRUFF database.

**(b)** Distribution of the FWHM values in Raman spectra from the RRUFF database.

**Figure 4.4:** Distribution of FWHM values in measured signals. The dataset has been selectively filtered to omit signals predominantly consisting of noise or those characterized by extensive peak overlap.

Consequently, in-depth analysis of the RRUFF database allows for drawing the following conclusions regarding the properties of measured XRD patterns and Raman spectra:

- The signals vary in length, containing between 600 and 9000 measurement steps.
- The measured intensities display a considerable range, varying from a magnitude of one hundred to one million.
- The number of peaks can vary greatly in the scans.
- Considering only individual peaks in those signals, the FWHMs are mostly in the range of 2 to 30, with occasional exceptions reaching up to 60.

Therefore, when designing a synthetic dataset to determine an optimized neural network structure for XRD scans and Raman spectra, careful consideration of these characteristics observed in measured signals is essential.

### 4.2.3 Synthetic Benchmark Generation

Having explored the inherent characteristics of measured signals, this section shifts focus to the design of a synthetic benchmark dataset. This dataset is specifically tailored to evaluate different configurations of CNN architectures, which employ numerous filters that slide across the input for feature detection. Thus, while the measured signals exhibit varying lengths and peak counts, the specific peak count and signal length become secondary considerations for the synthetic dataset. Instead, the focus is on evaluating the CNN's ability to identify individual peaks in limited signals. The mechanics of the convolutional layers suggest that once an optimized structure is established for accurately detecting individual peaks, the same set of filters can be seamlessly shifted across more complex signals with longer lengths and larger peak counts to extract the same information.

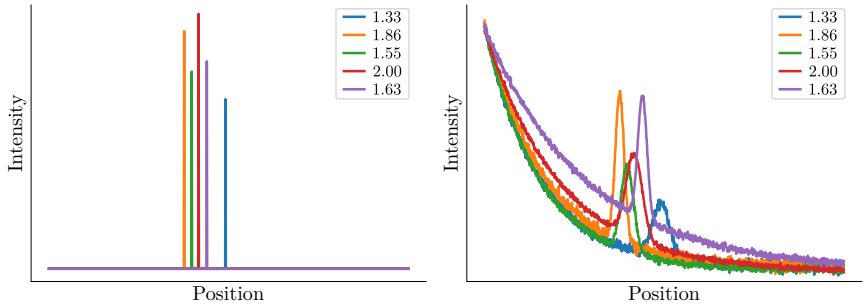
Moreover, it is advisable to limit the intensities of signals utilized for training or evaluating neural networks, as large counts facilitate overfitting [66]. Therefore, instead of evaluating the networks for analysis of signals with heights in the range of one thousand to one million, peaks should have much lower intensities. Furthermore, in signal analysis, the relevant information is generally not the absolute height of the peak but rather the encoded area beneath it. Hence, these synthetic peaks should reflect the same variation of peak widths found in measured signals. As the network is expected to perform effectively across a diverse range of signals, it is imperative to evaluate the full spectrum of FWHMs.

Finally, the synthetic signals are expected to contain background intensities and noise, which are commonly found in the measured scans. Such artifacts typically impact the measured intensity values and therefore complicate the accurate detection of peaks. As a result, a synthetic dataset is constructed that presents a straightforward task to evaluate the effectiveness of distinct neural network configurations: predicting the characteristic value of peaks in a noisy signal that is encoded in the area under the curve. To introduce an additional layer of complexity, the position of the peak varies, facilitating a thorough assessment of the network's utilization of shifting filters.

In summary, the synthetic dataset is designed according to the following principles:

- Signals consist of 1000 data points each.
- Each signal contains a peak with a height ranging from 0.4 to 2.5.
- The peak is characterized by a broad shape with an FWHM ranging from 2 to 60.
- The center of the peak varies between positions 350 and 650.
- Noise and background intensities are introduced in the signals.

Accordingly, Figure 4.5 showcases examples from the synthetic dataset. In Figure 4.5a, the ideal representation of the peaks is displayed, which highlights the variance of heights and positions. Furthermore, Figure 4.5b illustrates the appearance of the signals that are provided to the neural network models, which contain the typical peak shapes, background, and noise. This comprehensive approach ensures a realistic and challenging test dataset for assessing the performance of CNNs in analyzing signals with varying peak characteristics.




**(a)** Ideal representation of the peaks with characteristic heights and varying positions.

**(b)** Addition of background, noise and typical peak shapes.

**Figure 4.5:** Synthetic dataset for evaluating different neural network architectures. The goal is to predict the characteristic height of the peak, which is encoded in the area under the curve.

### 4.2.4 Evaluation of Distinct Network Configurations

To evaluate the effectiveness of various model configurations, each is tested with the task of predicting the exact value for the area under the curve. This evaluates the model's ability not only to identify different peak widths but also to handle noise and background interferences that affect peak characteristics. As a metric to quantify the accuracy of the prediction, the Root Mean Squared Error (RMSE) is used. In this particular evaluation, the synthetic dataset contains 10,000 signals which are split into 8,000 training and 2,000 validation samples. Each model undergoes training 11 times with random initializations, and the median RMSE is reported on the validation set.

The evaluation first involves a CNN with a single convolutional layer to explore the effect of various kernel sizes and filter counts. This basic network includes a flattening operation and a fully-connected layer with only one neuron that provides the prediction. Further tests include the addition of multiple subsequent fully-connected layers and the integration of pooling layers. This also includes the evaluation of the VGG, ResNet, and Inception architecture. Finally, regularization techniques are examined, along with the exploration of using several fully-connected layers prior to the output.

#### Single Convolutional Layer

Initially, the analysis is focused on CNNs with a single convolutional layer. These CNNs were tested with varying configurations, including kernel sizes of 17, 51, 127, 251, and 517, and filter counts of 8, 32, 128, 256, and 512. This methodical testing aims to identify the optimal combination of kernel size and filter count for accurately capturing the essential characteristics of the peaks.

Consideration of kernel size can be further motivated in the context of the synthetic dataset's FWHMs, which range from 2 to 60. For instance, a kernel size of 17 captures only a limited portion of the peak. Conversely, a kernel size of 51 closely aligns with the FWHM of peaks in the dataset, effectively covering most of the

peak area and facilitating the identification of features based on the peak's slopes. Furthermore, a kernel size of 127 (and greater) is large enough to accommodate most peaks in the synthetic dataset within a single kernel, thereby providing a comprehensive view of each peak's characteristics.

Table 4.1 showcases the performance metrics of various network configurations. Notably, networks with more filters and larger kernels performed better than those with fewer parameters. In particular, a filter count of 8 results in worse RMSE values for all kernel sizes tested here. Additionally, networks with kernel sizes of 17 and 51 consistently performed worse than those with kernel sizes of 127 or larger. This suggests that the kernel size should be large enough to accommodate the full range of intensities in a single kernel, rather than only half the width.

Although the network with 512 filters and 251 kernel size achieved the best overall Root Mean Squared Error (RMSE) of 1.27%, the benefits of incorporating additional filters or larger kernels are substantially reduced beyond a certain threshold. For example, the CNN with a kernel size of 127 and 128 filters achieved an average RMSE of 1.37%, only slightly larger than the best overall configuration. However, this network has fewer parameters to adapt, making it more efficient in terms of training and application. Therefore, it raises the question of whether networks with filter counts and kernel sizes surpassing this configuration are truly essential.

**Table 4.1:** Median RMSE scores for CNNs with a single convolutional layer on the single peak dataset. The best-performing configuration is highlighted in bold formatting.

| | | Kernel Size | | | | |
|---|---|---|---|---|---|---|
| | | 17 | 51 | 127 | 251 | 517 |
| No. of filters | 8 | 3.63% | 3.30% | 2.08% | - | - |
| | 32 | 3.28% | 2.71% | 1.57% | 1.39% | 1.39% |
| | 128 | 3.09% | 2.54% | 1.37% | 1.29% | 1.32% |
| | 256 | - | - | 1.32% | 1.27% | 1.29% |
| | 512 | - | - | 1.31% | 1.27% | 1.29% |

### Multiple Convolutional Layers

In practical applications, however, CNNs are not limited to the use of a single convolutional layer. Consequently, the ensuing analysis evaluates how incorporating additional convolutional layers can further enhance network performance on the single peak identification task. To this end, configurations that previously exhibited enhanced performance, particularly those with a high count of filters and a kernel size of 127, are retained. However, these networks are now modified to include more convolutional layers.

Figure 4.6 presents a visual comparison of the performance across various neural network architectures. The networks employing convolutional layers vary in complexity, ranging from one to eight layers, as denoted by the labels "Conv-1" through "Conv-8". Here, the grey boxes illustrate the range of RMSE values recorded for each architecture, encompassing the minimum and maximum RMSE observed. The median value of these RMSE values is distinctly marked with a red line within each box, providing a clear visual indicator of the central tendency in the performance of each network architecture.

Generally, a trend is observed where the RMSE metric decreases as the number of convolutional layers in the networks increases. For instance, the CNN with a single convolutional layer recorded a median RMSE of 1.37%. In contrast, this error metric was reduced to 1.11% in the model utilizing six convolutional layers. However, the Conv-8 model deviates from this general trend regarding the number of layers and RMSE. Notably, its overall performance is inferior to that of the Conv-6 model despite stacking more layers. This observation suggests that the Conv-8 model memorized the training samples (overfitted) rather than learning generally applicable features.

Furthermore, the computational complexity of the network increases greatly with the addition of each convolutional layer. Considering an input configuration with 64 channels, a kernel size of 127, and 64 filters, every added layer contributes 524,288 parameters to the overall network. As a result, while the Conv-6 model achieves the best RMSE metric at 1.11%, it encompasses roughly six times the

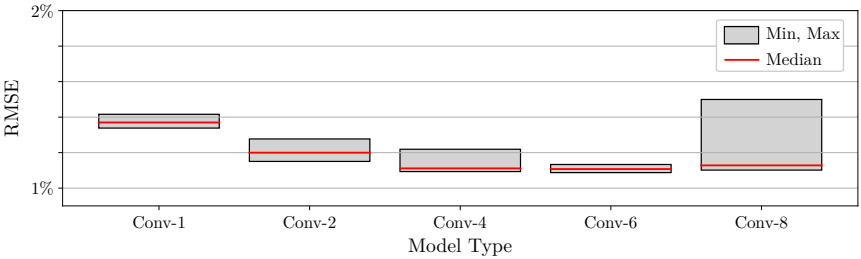



**Figure 4.6:** RMSE scores for different neural network architectures on the single peak signal dataset. The grey boxes indicate the best and worst performance across 11 models; the red line indicates the median. Networks using convolutional layers are specified with "Conv-$n$", where $n$ defines the number of convolutional layers in the network.

number of parameters found in the Conv-1 model, amounting to a total of over 3 million parameters. In contrast, the Conv-1 network demonstrates a relatively close performance to Conv-6, with only a 0.26 percentage point difference in RMSE. Therefore, it is evident that the enhancement in prediction accuracy comes at a significant cost, with only a slight improvement in the resulting metric.

### Pooling

In the field of image recognition, the computational cost associated with large kernel sizes is often mitigated by incorporating pooling layers, which serve to reduce the dimensionality of the input data. The concept of a receptive field refers to the spatial extent to which a convolutional kernel can gather information [66]. When a pooling layer downscales the data by a factor of two, the effective receptive field of a kernel of consistent size is doubled. Consequently, the subsequent investigation involves the use of pooling layers to utilize networks with smaller kernel sizes. This has the potential for a more efficient network design without compromising the network's ability to accurately identify the broad peaks.

For image recognition networks, consistent use of narrow kernel sizes across all layers is typical. Hence, one approach to reducing the size of the peak identification networks is to use a smaller kernel size (e.g., 51 instead of 127) in combination with pooling layers throughout the network. This approach,

**Table 4.2:** Median RMSE scores for CNNs incorporating pooling operations compared to the Conv-6 model. The labels A, B, and C correspond to distinct strategies for evolving kernel sizes.

| Network | Conv-6 | Pool-A | Pool-B | Pool-C |
|---|---|---|---|---|
| Kernels | 127 (6) | 51 (5) | 127-63-31-15-7 | 127-63-42-31-25 |
| RMSE | 1.11% | 1.20% | 1.22% | 1.18% |
| Parameters | 2,748,331 | 855,296 | 499,712 | 684,032 |

however, necessitates that the narrow convolutional filters in the initial layers retain the relevant information for subsequent layers with a larger receptive field. If a large kernel size in the initial layers is required for effective identification of the distinct peaks, an alternative strategy involves progressively reducing the size of the convolutional kernels. Consequently, additional evaluation is essential to validate the use of pooling layers and to determine an optimal configuration of convolutional kernel sizes that balances model performance and size.

Hence, the networks that integrate pooling layers are evaluated on the synthetic dataset and compared to the performance of the Conv-6 model. Given that the input has a length of 1000, introducing six pooling layers would yield an excessively narrow feature vector, shorter than the convolutional kernels. Consequently, the subsequent networks utilize only five convolutional layers to address this constraint. The initial pooling network adopts a consistent kernel size of 51 for all convolutional layers ("Pool-A"). In contrast, the second pooling model starts with a kernel size of 127 and then decreases the kernels by a factor of 2 for each pooling layer ("Pool-B"). Thus, the receptive field for the convolutional kernels in this network remains consistent. The final pooling model also starts with a 127 kernel size but implements a less aggressive reduction ("Pool-C"). While this also reduces the size of the neural network model, it concurrently enables the networks to process even larger features, given the increase of the receptive fields in the later layers.

Table 4.2 displays the median RMSE scores alongside the number of parameters for each architecture. Notably, the use of pooling layers greatly reduces the number of parameters in the network, while conserving the level of performance.

All three approaches (A, B, C) yield nearly identical RMSE scores; however, the third strategy emerges as the most successful. This suggests that employing a large kernel size in the initial layers could be beneficial for achieving optimal performance metrics. However, it is crucial to highlight that this evaluation specifically considers architectures where convolutional and pooling layers are sequentially stacked. There exist more advanced architectures in the domain of image recognition that have the potential to significantly enhance the peak identification capabilities of the models.

### Advanced Architectures

Therefore, neural networks that resemble the VGG, ResNet, and Inception architecture are also tested on the synthetic dataset. The performance metrics for these architectures, in terms of RMSE scores and parameter counts, are detailed in Table 4.3. The VGG architecture distinguishes itself by stacking multiple convolutional layers between the pooling operations in the network. Despite this arrangement, VGG's performance does not surpass that of the Conv-6 model, which does not utilize pooling. In a similar vein, the ResNet model, known for its intricate architecture, incorporates a greater number of convolutional layers compared to the other networks evaluated. However, it also does not achieve a better RMSE metric than the Conv-6 model.

The Inception network is designed to use parallel convolutional layers, each with varying kernel sizes, to capture features of different scales. However, previous investigations have highlighted that larger kernel sizes are particularly crucial for accurately detecting characteristic peak shapes in the data. Consequently, when evaluating the Inception network, it is observed that its performance does not match that of other network architectures. This outcome suggests that the Inception network, despite its notably lower parameter count, does not offer a significant advantage in this specific application.

Therefore, networks with more parameters do not necessarily perform better than smaller models, e.g., the Pool-C model, on the synthetic dataset. This lack of

**Table 4.3:** Median RMSE scores for CNNs with different architectures for stacking convolutional layers. "VGG" specifies a network with VGG-like architecture, "ResNet" for a ResNet-type model, and "Inception" for an Inception-like CNN. All networks have 64 filters.

| Network | Conv-6 | Pool-C | VGG | ResNet | Inception |
|---|---|---|---|---|---|
| RMSE | 1.11% | 1.18% | 1.12% | 1.16% | 1.25% |
| Parameters | 2,748,331 | 684,032 | 2,616,513 | 6,273,601 | 379,713 |

improvement could be attributed to overfitting, as large models are potentially over-parametrized and memorize the training samples, limiting their generalization ability. To validate this hypothesis, the use of regularization methods has to be evaluated.

### Regularization

Various types and configurations of regularization have to be considered. To this end, the Conv-6, VGG, and Pool-C models are complemented with different regularization methods during their training phase. One such modification includes the integration of batch normalization layers at various points throughout the network. Additionally, the application of dropout regularization is explored in two distinct configurations. The first approach involves implementing dropout regularization after each convolutional layer, in the following referred to as "convolutional dropout". The second approach applies dropout regularization only just before the final output layer of the network. Each of these configurations is tested separately to determine how they influence the model's ability to perform the regression task accurately, thereby providing insights into the optimal use of regularization techniques in neural network training for signal analysis.

Figure 4.7 displays the RMSE scores across various regularization configurations, providing context by including the performances of networks without any added regularizations for comparison. To emphasize the differences in performance, a logarithmic scale is employed. Notably, the addition of convolutional dropout resulted in the worst performance metric in this study, exceeding 13% RMSE. While

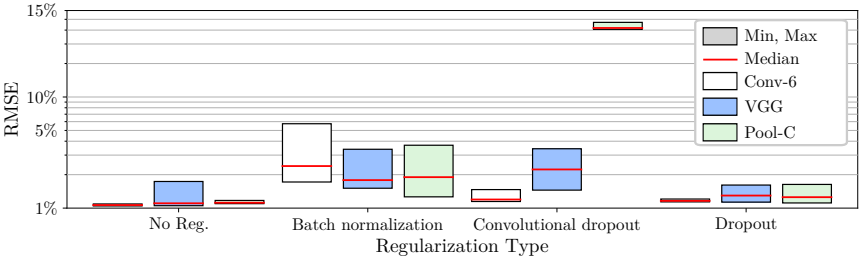



**Figure 4.7:** Performance of various CNNs with different regularization techniques.

convolutional dropout did not affect the Conv-6 and VGG models to the same extent as Pool-C, the performance of all models is worse than their counterparts without regularization. Similar effects are observed with batch normalization, resulting in inferior performance across all models. Only dropout regularization had a minimal impact on model performance, with the model achieving approximately the same metrics with and without this regularization technique. However, this evaluation shows that the performance of the large models cannot be further improved by adding regularization techniques.

### Parameter Count

While various model configurations have been tested, the primary focus has predominantly been on the overall prediction accuracy measured by RMSE. However, in order to pick an optimized neural network architecture for use within the novel material identification framework, it is imperative to also consider the size of the model. Opting for a network with fewer parameters not only reduces computational demands during training and application but also diminishes the likelihood of overfitting. The evaluation of different regularization methods reveals that employing batch normalization or dropout does not effectively prevent overfitting in larger models, underscoring the advantage of utilizing models with fewer parameters to mitigate this issue.

Accordingly, Figure 4.8 presents an exhaustive overview that compares the parameter count and the median RMSE scores across different network architectures.

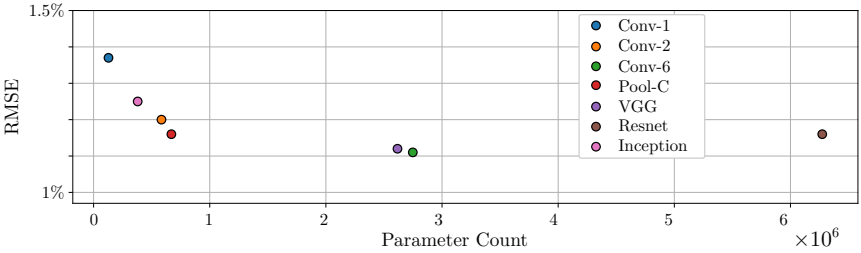



**Figure 4.8:** Contrasting the median RMSE scores and the number of parameters in different networks for the single peak signal dataset. While the RMSE decreases as the parameter count increases, no network configuration achieved an error score below 1%.

This visual representation clearly demonstrates the utility of convolutional layers in extracting peak information from signals. Notably, this reveals a general trend where the prediction error tends to decrease as the number of parameters in the models increases. However, no network achieves an RMSE metric lower than 1.1%, and beyond the Pool-C model, which has about 600,000 parameters, the benefit of adding more parameters is minimal. This observation is crucial in understanding the balance between network complexity and performance efficiency in the context of extracting peak information from noisy signals.

### 4.2.5 Proposed Network Structure

Contrasting previous studies, which often focused on introducing novel model architectures tailored to their particular dataset, this thesis aims to develop a unified neural network structure that achieves high performance metrics for all types of Raman spectra and XRD patterns. Therefore, the systematic evaluation of the synthetic datasets, designed to mirror the properties found in measured signals, has facilitated a comprehensive comparison of various neural network architectures. This in-depth analysis has not only provided insights into the relationship between network parameters and signal properties, such as the connection between kernel size and peak FWHMs, but it has also revealed that certain regularization techniques commonly used in established networks can negatively affect performance.

Accordingly, the previously defined open questions can be answered based on the results obtained through the evaluation of network configurations using the synthetic datasets. Approximately 64 filters are required to capture the varying appearance of peak shapes. In addition, a convolutional kernel size of 127 is required to accurately identify peaks in noisy signals with FWHMs of up to 60 data points. However, by using pooling layers, the kernel size can be reduced in later layers without affecting the performance of the model. To accurately identify individual peaks, multiple convolutional layers should be stacked, but more advanced concepts including VGG, ResNet, and Inception do not improve the performance of such peak identification networks. Furthermore, the dropout regularization technique is not helpful in these networks, and adding batch normalization or dropout between convolutional layers seriously degrades performance.

However, the analysis of the synthetic dataset only evaluates the ability of the networks to identify individual peaks in the signals. Measured signals typically contain multiple peaks, often overlapping, and the signals exhibit diverse lengths, as described in Section 4.2.2. While the convolutional kernels adeptly navigate signals of varying lengths, they prove most effective in identifying local features. Since the convolutional kernels glide over the input, the analysis of signals of different lengths is not a challenge, but these filters are only useful for identifying local features. To enhance the network's proficiency in identifying global features, the neural network structure should be extended with a fully-connected layer[1].

Therefore, a unified neural network architecture is proposed based on the results of the synthetic benchmark evaluation. Figure 4.9 illustrates the structure of the CNN, which resembles the Pool-C model. Therefore, the network consists of five convolutional layers that are accompanied by pooling operations. The input layer of the network can be adapted to match the size of the inputs, but it should include at least 600 neurons to ensure that the data is larger than the size of the convolutional layers after the pooling operations. After the convolutional stage,

[1] The addition of a fully-connected layer was also tested for the CNNs in the synthetic dataset but did not prove to be useful for identifying the individual peaks in the signal.

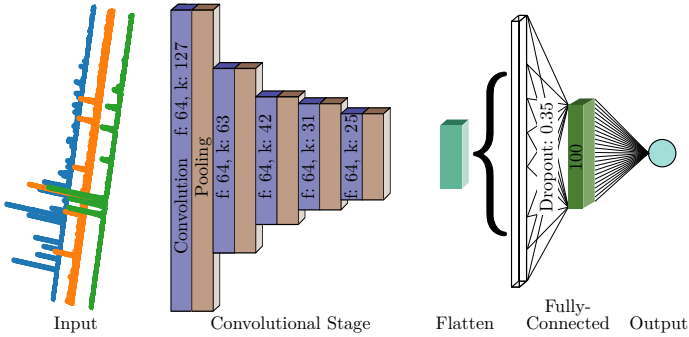



**Figure 4.9:** Proposed neural network architecture for use with spectroscopic signals and diffraction patterns. The network is split into several key components: the convolutional stage, the flattening, the fully-connected layers, and the output. f specifies the number of filters, and k the kernel size in the convolutional layers. The maximum pooling layers each half the length of the input.

the flattening operation reshapes the encoded features, and one fully-connected layer with 100 neurons is employed to detect the spatially distant features.

Furthermore, 35% of the connections between the flattened feature vector and the hidden layer are randomly dropped during training to mitigate overfitting. This regularization technique was the only one that did not substantially affect the network's performance. Thus, incorporating this regularization method ensures that the model does not risk overfitting, irrespective of the size and nature of the data used for training the neural network. This represents another measure to ensure the straightforward application of the model to datasets that analyze diverse material systems, affirming its robustness in accommodating a wide range of data variations.

Although the network is designed for use with both spectra and diffraction patterns, its application has thus far been exclusively demonstrated on signals containing only a single peak. Additionally, despite testing various neural network configurations, it remains uncertain whether an alternative architecture presented in a previous study might outperform the proposed network in analyzing such

fingerprints. Consequently, further tests are imperative to precisely assess the performance of this model when applied to measured signals encompassing multiple peaks.

## 4.3 Comparative Analysis of Network Structures

### 4.3.1 Configurations of Established Networks

Numerous alternative neural network structures have been presented for the analysis of Raman spectra and diffraction patterns. The first documented applications of convolutional neural networks in the field of one-dimensional pattern analysis were presented by Park et al. [42] and Liu et al. [17], with each utilizing a network comprising three convolutional layers. Furthermore, there is a broad spectrum of distinct network configurations, employing varying numbers of filters, kernel sizes and convolutional layers. Moreover, some of these networks utilize more advanced network architectures, such as VGG, ResNet, and Inception [15, 16, 31], and integrate alternative regularization methods including batch normalization and dropout between the convolutional layers.

Table 4.4 provides a summary of the various network architectures that have been presented in previous studies, in addition to the proposed model. Most architectures incorporate multiple convolutional layers, except for the network developed by Mozaffari and Tay [71]. The proposed network architecture distinguishes itself from alternative configurations due to its large kernel size. In addition, only three network configurations have fewer parameters than the proposed architecture. Although the network by Mozaffari and Tay features a single convolutional layer and fewer fully-connected neurons than the novel model, it still contains more parameters overall. These additional parameters are predominantly part of the fully-connected layer, necessary for mapping the intensities encoded in the feature vector to the output. This highlights the benefits of using pooling layers,

**Table 4.4:** Configurations of distinct neural networks for analyzing XRD patterns and Raman spectra. The number of filters is reported for the last convolutional layer (indicative of the feature vector), and the kernel size for the first convolutional layer. The regularization methods dropout and batch normalization are abbreviated as DO and BN, respectively. The number of layers is determined exclusively by considering the convolutional layers. To determine the model size, an input size with 5,000 data points was considered.

| Network | No. Layers | Filters | Kernel | Reg. | FC-Neurons | Parameters |
|---|---|---|---|---|---|---|
| Park et al. [42] | 3 | 80 | 100 | DO | 770 | 3,170,081 |
| Liu et al. [17] | 3 | 64 | 21 | BN, DO | 2048 | 81,940,865 |
| Oviedo et al. [44] | 3 | 32 | 8 | - | - | 8,577 |
| Fan et al. [45] | 2 | 64 | 5 | DO | 4608 | 290,406,145 |
| Vecsei et al. [46] | 3 | 80 | 100 | DO | 3450 | 11,785,841 |
| Ho et al. [31] | 26 | 100 | 5 | BN | - | 1,261,941 |
| Wang et al. [15] | 7 | 64 | 5 | DO | 390 | 2,462,861 |
| Lee et al. [16] | 2-9 | 64/332 | 50/20 | DO | 4000 | 186,617,139 |
| Szymanski et al. [43] | 6 | 64 | 35 | DO | 4300 | 19,612,925 |
| Schuetzke et al. [68] | 3 | 64 | 20 | DO | 2500 | 24,848,373 |
| Sang et al. [70] | 16 | 512 | 3 | BN, DO | 512 | 47,575,617 |
| Bhattacharya et al. [72] | 4 | 32 | 2 | DO | 1000 | 305,037 |
| Mozaffari and Tay [71] | 1 | 32 | 3 | - | 64 | 10,240,257 |
| *Proposed* | 5 | 64 | 127 | DO | 100 | 1,666,505 |

as the proposed model reduces the dimensionality of the input by a factor of 32, resulting in fewer connections for the fully-connected neurons.

Nonetheless, the different neural network architectures described here each achieved high accuracy metrics for the classification of XRD patterns or Raman spectra in their respective publications. Hence, while the proposed network structure has been specifically designed for optimal performance when analyzing such signals, it is mandatory to compare its performance to that of the other models. However, as explained in Section 4.2.1, the absence of large-scale datasets with reliable labels complicates an extensive comparison. Thus, another synthetic dataset is introduced that presents a classification task involving the characteristic fingerprints of arbitrary substances in signals. As a result, this dataset provides an unprecedented opportunity to benchmark different network architectures.

### 4.3.2 Benchmark Dataset for Comparative Analysis

The synthetic benchmark dataset, as described in this section, allows for evaluating the accuracy and the computational efficiency of the neural network models. Unlike the initial synthetic dataset, which was designed to assess the networks' proficiency in identifying key properties of a single peak amidst noise, this additional dataset is intended to test the networks' ability to recognize a characteristic fingerprint. Accordingly, the dataset includes signals that either display this specific fingerprint or contain different patterns, including instances where the fingerprint is present alongside additional peaks. The objective for the neural networks in this scenario is to determine a single outcome: whether the input signal exclusively matches the predefined fingerprint or not.

While the samples of this dataset are generated artificially, the signals are intended to resemble measured scans. Consequently, properties of the signals from the RRUFF database are systematically incorporated into the design of the synthetic dataset. This approach ensures that the results obtained from evaluating the networks on this synthetic data set are expected to reflect the model's performance when applied to the analysis of real measured signals.

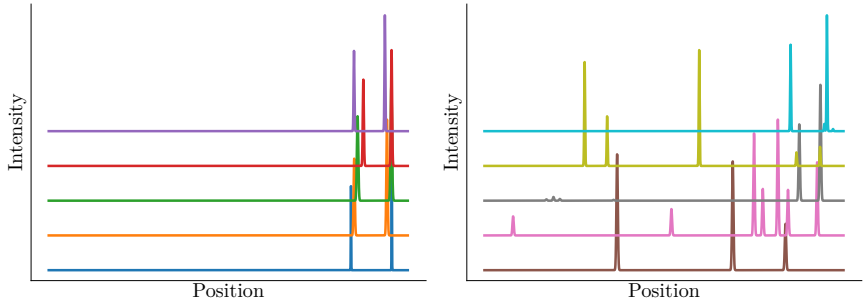
Therefore, the dataset comprises signals of length 5000, aligning with the median signal length observed in the RRUFF database. Furthermore, a unique fingerprint is defined which contains multiple peaks with characteristic positions and heights. The peaks are represented by broad shapes with FWHMs between 2 and 30 data points[2] and the signals contain noise and background intensities. Furthermore, the positions, intensities, and widths of the fingerprint can vary, providing a realistic representation of the diversity typically observed in actual signals. Details

[2] Signals from the RRUFF database contain peaks with FWHMs of up to 60. However, the majority of established network architectures utilize smaller kernel sizes, indicating that these networks may not have been specifically tailored for broad peaks. Given that 99% of the signals in the RRUFF dataset have peak widths less than 30 FWHM, a more limited scenario is examined here. Nevertheless, it is essential to note that the proposed architecture is designed to analyze signals with FWHMs of up to 60.

regarding the degree of variation are explained in Section 5.3 and Appendix Section A.4.2.

In addition, signals are also generated that serve as negative examples for training and evaluating the various neural network configurations. These alternative patterns feature peaks with positions and heights that differ distinctly from the target fingerprint, thereby providing a contrasting set of data. Additionally, some of the generated signals include patterns that closely resemble the key fingerprint but are distinguished by the presence of additional peaks. In line with the characteristics observed in the measured signals from the RRUFF database, the peaks within the characteristic fingerprint may overlap with other peaks. This aspect is deliberately incorporated into the dataset to assess the neural networks' robustness and accuracy in scenarios where peak overlap occurs.

Accordingly, Figure 4.10 presents signals from this benchmark dataset that depict signals without noise and background. Figure 4.10a offers a visual representation of this characteristic pattern that the neural networks are tasked to identify within the signals. The illustrated examples demonstrate the actual variations in the fingerprint, such as shifts in peak positions and changes in peak intensities and widths. Furthermore, Figure 4.10b presents examples, which the neural networks are tasked to differentiate from the actual samples containing the fingerprint. The signals illustrated in brown and pink exhibit patterns that are distinctly different from the target fingerprint. In contrast, the grey, olive-green, and cyan plots represent negative examples where the fingerprint is present alongside additional peaks. Notably, in the cyan plot, these additional peaks partially overlap with one of the target fingerprint peaks and are difficult to identify due to the small peak heights. This scenario poses a significant challenge for the models, testing their ability to accurately identify the fingerprint despite the presence of overlapping and nonessential peaks.




**(a)** Positive examples of the signals in the synthetic fingerprint identification dataset. The characteristic pattern is varied with respect to the peak positions, heights, and widths.

**(b)** Negative examples of the signals in the synthetic fingerprint identification dataset. The patterns are either completely random or show the fingerprint to identify amidst complementary peaks.

**Figure 4.10:** Synthetic Fingerprint Identification Dataset: This dataset is specifically crafted to assess the capability of various neural network architectures in recognizing a unique fingerprint. It mirrors properties typically found in measured scans, thereby ensuring the relevance of the dataset to real-world scenarios.

### 4.3.3 Results of Comparative Analysis

In total, the models are evaluated using the synthetic benchmark dataset, which comprises 10,000 signals. Among these, 5,000 depict the variation of the fingerprint without additional peaks (referred to as "positive examples"), while the remaining 5,000 represent alternative patterns, including cases where the target fingerprint is present among impurities (considered "negative examples"). Therefore, each network architecture is specifically designed with a single output node, employing the sigmoid activation function. The training of these networks is conducted using the binary cross-entropy loss function. To ensure a robust evaluation, the dataset is divided into training and validation sets, with an 80%-20% split. Thus, the quality of the prediction is measured using the accuracy metric calculated on the validation set, and the computational complexity is estimated based on the number of parameters in each model.

Table 4.5 presents the outcomes of the network benchmark, organized by model size, and highlights the highest accuracy score. Generally, most networks achieved high classification metrics, with only three models falling short of achieving an

accuracy exceeding 90%. Moreover, the performance gap between the smallest and largest model is less than 4 percentage points, indicating that additional parameters may not necessarily enhance model performance. The proposed model, with only 1.7 million parameters, achieved the best metric, accurately classifying 99% of the 2,000 validation samples. While a few models approached this performance, with four models achieving a score of 98.5%, the small numerical difference corresponds to another 10 misclassified signals.

The proposed model achieved the highest accuracy score while employing less than 1% of the parameters compared to the largest model. This underscores the effectiveness of designing the model using the synthetic dataset containing signals with individual peaks. Despite utilizing this limited dataset, the performance of the proposed model extends well to identifying signals with multiple peaks. This emphasizes the functionality of the convolutional filters, which slide across the input of varying lengths and identify position-irrelevant features. Therefore, given the model's demonstrated performance on the benchmark dataset inspired by measured signals, it is reasonable to expect that its effectiveness will translate well to the analysis of measured scans.

## 4.4 Application of Network Model

The presented neural network configuration, as illustrated in Figure 4.9, has been precisely optimized for a dataset that mirrors the typical parameters commonly observed in measured scans. This optimization specifically accounts for peaks with FWHM values ranging between 2 and 60. As a result of this optimized approach to determining adequate network configurations, the presented model is suitably equipped to handle most data encountered in practical applications. In addition, the model can be flexibly applied to signals of varying lengths because the convolutional filters identify all relevant peaks in the measurements by sliding over the input. This adaptability implies that the network can be effectively applied to a wide range of data without necessitating any significant modifications to its

**Table 4.5:** Results of the network evaluation on the synthetic fingerprint benchmark. The networks are sorted in ascending order by the number of parameters. *Two networks that employ batch normalization classified all samples of the validation set as "positive".

| Network | Accuracy | Parameters |
|---|---|---|
| Oviedo et al. [44] | 91.78% | 8,577 |
| Bhattacharya et al. [72] | 72.76% | 305,037 |
| Ho et al. [31] | 50%* | 1,261,941 |
| *Proposed* | **99.01%** | 1,666,505 |
| Wang et al. [15] | 98.05% | 2,462,861 |
| Park et al. [42] | 98.53% | 3,170,081 |
| Mozaffari and Tay [71] | 94.31% | 10,240,257 |
| Vecsei et al. [46] | 98.53% | 11,785,841 |
| Szymanski et al. [43] | 98.53% | 19,612,925 |
| Sang et al. [70] | 97.57% | 47,575,617 |
| Liu et al. [17] | 50%* | 81,940,865 |
| Lee et al. [16] | 98.53% | 186,617,139 |
| Fan et al. [45] | 95.26% | 290,406,145 |

underlying architecture, thereby offering a versatile tool for analyzing measured scans with varying peak characteristics.

However, the application to signals with peaks exceeding the FWHM of 60 measurement points has not been tested. Broad peaks of this magnitude could be the consequence of a measurement with an extremely small step size. In such cases, the signals can be resampled to larger step sizes to ensure that the resulting peaks have widths below 60 FWHM. Nonetheless, researchers would be aware if they perform measurements with such extraordinary settings, so corresponding procedures can be applied before feeding the signals into the neural network.

Hence, the training and application procedure of the neural network model is as follows: The training process begins by providing the neural network with training data that represents the patterns of the material system being analyzed and defines the measurement range and step size of the signals. Then, the neural network

model, including all its incorporated layers, is adapted to the length of the training data. Through this training, the network learns to identify unique fingerprints based on relevant features encoded in the feature vector. Thus, once the model is trained, it is critical to ensure that only data with appropriate measurement ranges and step sizes are analyzed. This ensures optimal performance because the network is tuned to detect patterns within the specified measurement steps.

# 5 Implementation

## 5.1 Overview

The preceding chapters thoroughly outlined the conceptual framework for automated analysis of XRD scans and Raman spectra using neural networks. This included a detailed description of the method for simulating training data and the complexities of determining an adequate network architecture. The focus now shifts to the practical implementation of these approaches. Accordingly, this chapter is dedicated to explaining the step-by-step process of implementing the algorithms and approaches previously discussed, detailing the technical aspects and challenges encountered during the implementation phase. Several software packages have been developed as a key component of this thesis, containing the implementation of various elements utilized in the previous chapters. This chapter offers a technical description of the algorithms and explains how these software packages can effectively integrate the presented concept into existing material discovery platforms.

Before addressing the specific details of the implementation, it is imperative to outline the essential requirements for software development in the context of this thesis. Firstly, the aspect of simulating training data forms the backbone of the neural network's learning process. This approach incorporates existing simulation tools, so seamless integration is crucial to enable the simulation of training signals. Secondly, training and evaluating the neural networks is an essential element, including generating synthetic datasets to analyze distinct network configurations. Lastly, the seamless processing of acquired signals, generated by

measuring the samples produced by the material discovery platform, is a fundamental requirement. These measurements are typically provided as text files containing numerical values describing the measurement steps and the corresponding measured intensities. Consequently, selecting a development environment and programming language that are precisely aligned with these defined requirements becomes essential.

Given that neural networks constitute the central element of the data analysis concept, as presented in Chapter 2, the Python programming language emerges as an optimal choice for implementing the essential algorithms. This decision is guided by Python's widespread adoption within the deep learning community, notably in developing libraries such as *TensorFlow* [88] and *PyTorch* [89]. Thus, implementing the data simulation approaches in the same language ensures seamless integration when using the generated data in the neural network training routines. Moreover, Python provides a rich ecosystem featuring established libraries like *NumPy* and *SciPy*, which mainly facilitate the processing of vectors and matrices. Such capabilities are instrumental in efficiently processing and storing the simulated training data but also in importing and preprocessing the experimental signals.

In the context of this thesis, several software packages have been developed. Figure 5.1 provides a graphical overview of the packages that have been developed in the context of this thesis. This includes methods for accurate simulation of powder diffraction scans (*python-powder-diffraction*), as well as the comparison of distinct neural network models using a synthetic benchmark (*spectra-network-benchmark*). Most notably, the central concept of this thesis, the identification of target materials in produced samples based on their characteristic fingerprint, is implemented and available for straightforward integration into existing material discovery platforms in the form of the *crystal-id* package. To accomplish this, the *crystal-id* package includes an optimized neural network structure that has been developed using the *spectra-network-benchmark* repository. Furthermore, optimized routines for the simulation of a large-scale dataset, which have originally been developed for the *python-powder-diffraction* library, are integrated into the *crystal-id* package to enable the training of the neural network model.

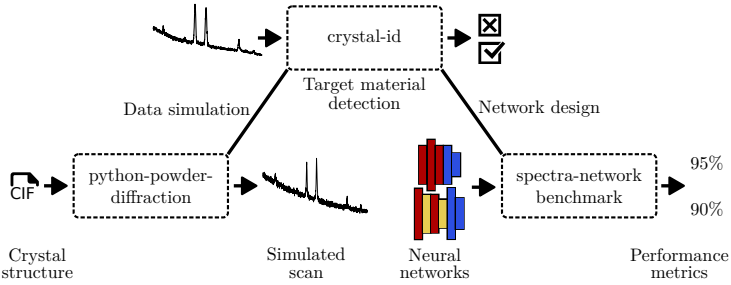



**Figure 5.1:** Overview of the software packages that have been developed as part of this thesis. The *python-powder-diffraction* library provides functionality to simulate diffraction patterns for provided materials (in CIF format). The *spectra-network-benchmark* repository includes model implementations and scripts to assess the performance of distinct neural network architectures for the analysis of spectra-like signals. The *crystal-id* package contains the functionality to identify target materials using a neural network model. Therefore, the package includes data simulation methods and results of the network benchmark, which have been developed or obtained using the other packages.

Accordingly, this chapter is structured as follows: Section 5.2 provides the technical implementation to simulate realistic XRD patterns. Subsequently, Section 5.3 explains the algorithms and procedure for testing different network configurations using synthetic datasets. Finally, Section 5.4 introduces the technical implementation of the *crystal-id* framework. Furthermore, Section 5.4 explains how to integrate the framework for the analysis of material discovery data at hand.

## 5.2 XRD Powder Pattern Simulation

### 5.2.1 Package Overview

At the start of this dissertation, the training of neural networks for analyzing XRD patterns predominantly utilized data sourced directly from crystallographic databases. A notable example is the ICSD database [36], which provides access to a vast array of crystalline material entries and offers the download of simulated diffraction patterns for their database entries with specified radiation wavelengths, measurement ranges, and step widths. While these simulated patterns include the

representation of peaks as broad profiles, they lack the incorporation of additional experimental artifacts like background intensities and noise. Moreover, the limitation to database entries means only those materials listed can be simulated.

As an alternative to the closed-source simulation algorithm of the ICSD, the well-established *pymatgen* library [76] provides the functionality to calculate diffraction patterns using the Python programming language. *Pymatgen* contains methods to read and convert crystallographic information files (CIF) from various sources and is therefore not limited to the entries from the ICSD. The *XRDCalculator* object of *pymatgen* includes the functionality to calculate the diffraction position and intensities for a given structure, wavelength, and two-theta range. Nonetheless, the *pymatgen* modules do not consider the broad appearance of peaks in measured signals, and the package also does not include methods for simulating background intensities and noise.

Accordingly, the *python-powder-diffraction* package was developed. This approach takes a material's structural representation as input in the form of a CIF and simulates diverse effects that result in altered peak positions and heights in the corresponding diffraction pattern. Furthermore, the artificial patterns are complemented by background intensities and noise, which ensures that the generated signals are virtually indistinguishable from measured scans. Hence, it enables the generation of data that captures the variability of the characteristic fingerprint without requiring the availability of several database entries. The package is publicly available on GitHub at
`https://github.com/jschuetzke/python-powder-diffraction`.

Figure 5.2 provides an overview of the *python-powder-diffraction* library with the modules highlighted in yellow and the user-executable script in blue. While the *Powder* class and the *noise* function contain the implementations of the algorithms discussed in the previous chapters, the *generate training data* script can be executed by the user to generate a dataset of diffraction patterns (signals and the corresponding labels) from a list of CIFs. The following subsections are dedicated to a detailed explanation of the core modules and objects in the *python-powder-diffraction* library.

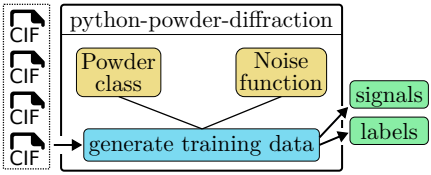



**Figure 5.2:** The structure of the *python-powder-diffraction* library. At its core, the *Powder* class and the *noise* function can be employed to simulate realistic powder diffraction patterns. Furthermore, the *generate training data* script offers a straightforward function to utilize the library for generating large-scale simulated datasets via the command-line interface.

### 5.2.2 The Powder Object

The core functionality of the *python-powder-diffraction* package is built around the *Powder* object. This class is based on the *pymatgen Structure* object and extends its features to produce varied diffraction patterns. The structure instance provides functionality to read the crystalline information from CIFs and represents the materials. Subsequently, the *Powder* class introduces methods to manipulate this information.

The position of peaks in a powder diffraction pattern is determined by the dimensions and symmetry of the crystal lattice, as explained in Chapter 1. To facilitate the variation of these peak positions, the presented software enables the manipulation of the unit cell parameters for a given material. This manipulation is achieved by randomly varying the unit cell parameters while adhering to the constraints imposed by the crystal system of the material. Thus, employing the *vary_strain* method within the *Powder* class makes generating multiple variations of the base crystal structure possible.

Algorithm 1 provides a detailed implementation of this methodology. The input for the algorithm is a $str_x$ structure with parameters $a, b, c, \alpha, \beta, \gamma$. Furthermore, the constraints of the crystal system are described by $c_x$. To randomly modify the lattice parameters, function $Random(m, n)$ is necessary to draw a value $r$ from distribution [a,b]. The maximum degree of the lattice variation is defined by $var_{\max}$, and the algorithm outputs the varied structure $str_y$.

**Algorithm 1** Generation of varied lattices in the *Powder* object.

**Require:** $str_x$ ▷ reference structure
**Require:** $var_{\text{max}}$ ▷ maximum variation
1: $a, b, c, \alpha, \beta, \gamma \leftarrow str_x$
2: $c_x \leftarrow str_x$
Get random factors for all lattice parameters
3: $r \leftarrow Random(1 - var_{\text{max}}, 1 + var_{\text{max}})$
Modify lattice parameters with random factors
4: $a, b, c, \alpha, \beta, \gamma \leftarrow (a, b, c, \alpha, \beta, \gamma) \cdot r$
Consider the constraints of the crystal system
5: $a, b, c, \alpha, \beta, \gamma \leftarrow c_x(a, b, c, \alpha, \beta, \gamma)$
Modified Structure
6: $str_y \leftarrow c_x(a, b, c, \alpha, \beta, \gamma)$

Furthermore, the *Powder* class integrates the *XRDCalculator* object to compute the diffraction patterns for the structures. Consequently, the *Powder* object takes the information regarding the wavelength as input and initializes a calculator to process the structures accordingly. Noteworthy, measured scans frequently show diffraction peaks from multiple wavelengths as a result of the K$\alpha_1$ and K$\alpha_2$ peaks in the X-ray profile that cannot be filtered out (see Section 1.3.2). Thus, the *Powder* object also takes multiple wavelengths as input and scales the computed diffraction peaks according to the ratio of the different peaks in the radiation profile. The functionality to obtain the computed position and intensities of peaks for the provided structure is integrated into the *get pattern* method of the *Powder* class.

In addition to the positional variations of peaks in diffraction patterns, the heights of these peaks are subject to variation due to a range of effects, one of which is the formation of preferred particle orientation in the powder (see Section 1.3.3). To simulate this phenomenon accurately, the *Powder* class has been designed to integrate the variation resulting from preferred orientation effects. For each computed diffraction position and intensity, the *XRDCalculator* provides information regarding the corresponding family of Miller planes. The variation process then

involves randomly selecting a Miller plane and a texture factor and then calculating the degree of orientation alignment of various planes in the pattern with this randomly chosen orientation. This process can lead to either amplification or reduction in peak intensities, depending on their alignment with the selected orientation. This functionality is implemented in the *vary texture* method of the *Powder* class.

While the diffraction pattern calculator accurately determines the exact positions of peaks in the pattern, the actual measured patterns yield information only at specific measurement steps. For the *Powder* object, it is necessary to define the $2\theta$ range and step size during the initialization of an instance. Consequently, aligning the computed peak positions with these measurement steps becomes imperative. This alignment is achieved through a straightforward procedure involving mapping the computed diffraction angles to their nearest measurement steps.

In the final stage of the pattern simulation process, the calculated intensities are transformed to resemble the broad profiles found in measured signals. This transformation is achieved by convoluting the discrete intensity values with Gaussian profiles of varying widths. The peak width in diffraction patterns is typically characterized using the Scherrer equation (Equation 1.3), which relates it to the grain size of the powder. In this approach, the parameters for the Gaussian profiles are established by randomly selecting the grain sizes of the powder. The random selection is performed by drawing from a uniform distribution, with a default range of 10 to 100 nm. These sizes are then used to determine the corresponding peak widths, which are further translated into the discrete Full Width at Half Maximum (FWHM) of the peaks, taking into account the step size attribute of the scans. To perform the convolution of the discrete signals with these profiles, the *SciPy* function *gaussian_filter1d* is integrated, which provides an optimized implementation of the convolution operation.

Hence, the *Powder* object takes the wavelength, $2\theta$ range, step size, and parameters regarding the pattern variation as input and generates varied patterns for the provided structure. Figure 5.3 provides a visual summary of the *Powder* class. Here, the attributes of the object are highlighted in yellow, the methods in gray

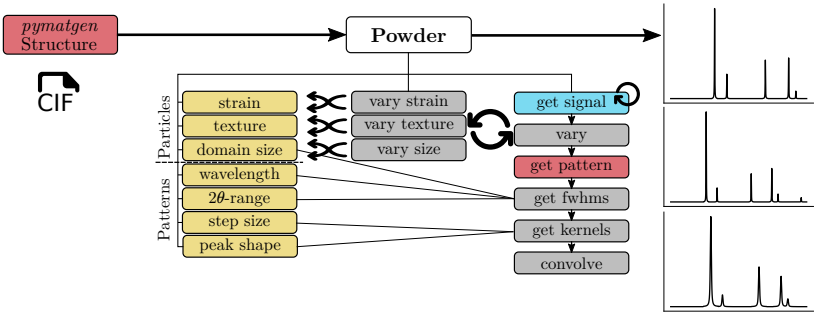



**Figure 5.3:** Simulation pipeline of *Powder* object in the *python-powder-diffraction* package.

(private) and blue (public), and the integration of the *pymatgen* functionality is depicted in red. The object takes a CIF as input, and the method *get signal* enables the unlimited simulation of varied patterns, which are subsequently saved in the *NumPy* format *npy* to ensure optimal saving and loading procedure for subsequent processing of the signals.

### 5.2.3 Noise Simulation

While the *Powder* class enables the generation of patterns with physics-informed position shift, height variation, and diverse peak shapes, the simulated patterns do not contain noise and baseline intensities that are typically found in measured data, as explained in Section 1.3.3. However, to generate realistic XRD patterns, it is mandatory to represent all artifacts present in the measured signals. Hence, the package's functionality must be extended to include such experimental artifacts in the simulated patterns.

In contrast to the variations to each pattern, which must be calculated individually, noise can be computed in bulk to optimize the data generation throughput. Thus, a noise generation algorithm is implemented that takes a batch of simulated signals as input and adds noise and background simultaneously for the different powder diffraction patterns. To simplify the implementation, the *NumPy* library includes

functionality to generate Gaussian noise for a given shape (e.g., 100 scans, 5000 data points each).

While noise is present in all measured scans, the level of noise, as described by the signal-to-noise ratio (SNR) can vary greatly between measurements. Typically, the maximum intensity attributable to the noise is approximately 20 times lower than the peak intensity of the highest diffraction peak [80]. The highest intensity in measured scans can vary substantially, depending on the instrument or acquisition times used, but simulated scans are typically scaled between 0 and 1 (see Section 3.2). Consequently, a straightforward approach is to draw random values from a Gaussian distribution with matching mean and standard deviation parameters that are subsequently added to the simulated patterns.

Hence, the implemented approach involves drawing random values from a Gaussian distribution with a mean of 0.5 and a standard deviation of 0.11, which ensures that 99.9% of the random values fall within the interval [0, 1]. Furthermore, any drawn values exceeding these limits are clipped, but this has no meaningful impact on the overall distribution of the random values. This approach results in noise values within the interval [0, 1], so the subsequent step involves scaling the noise according to the intended SNRs. For each scan, a unique noise factor is determined within the range from zero to the highest acceptable level of noise for simulated patterns. This factor is then multiplied with the generated noise to achieve the desired noise level and the generated noise is added to the simulated patterns. Hence, this approach guarantees the presence of unique noise in terms of exact values and SNRs for each simulated pattern, while preventing the occurrence of negative intensities.

Similarly, the background intensities can be simulated using the Chebyshev polynomials functionality provided by *NumPy*. Therefore, noise is generated in bulk for all simulated scans simultaneously, scaled, and added, together with the background, to the matrix of scans, which is magnitudes faster than looping over each scan and generating noise individually.

### 5.2.4 Package Utilization

While the *Powder* object and the noise simulation methods are elemental for generating varied signals for a single material, generating diffraction patterns for multiple phases is often necessary to create a comprehensive dataset. Consider the simulation of patterns for 500 unique phases, each with 100 variations. Table 5.1 presents the recorded times to initialize the *Powder* from a CIF file and generate 100 varied patterns on the development systems. In this implementation, the convolutional operation of representing the broadened peak shapes is the bottleneck of the signal generation approach. Hence, about 18.5 seconds are required to simulate 100 signals for a single phase.

Accordingly, simulating the patterns for 500 phases results in a simulation time of about 154 minutes. To enhance the throughput of this process, the data simulation script employs the *multiprocessing* library of Python. This allows the simulation task of distinct phases to be divided into separate processes distributed across the available processing cores, optimizing the use of computational resources. Thus, the same task can be performed in less than 20 minutes on a system with eight cores, assuming minimal overhead when integrating the *multiprocessing* modules.

Figure 5.4 provides a visual explanation of this procedure. Each core can select a different structure using the provided implementation and generate multiple variations through unique powder instances without affecting other computations. Accordingly, the speedup of this parallel approach is related to the number of cores

**Table 5.1:** Approximate times to utilize the Powder class for simulation of varied diffraction patterns.

| step | time [s] |
|---|---|
| initialization *Powder* instance | 1.5 |
| vary parameters | 0.04 |
| get signal | 0.13 |
| total (100 times) | 18.5 |

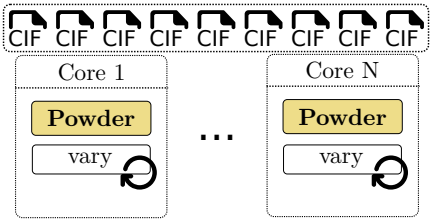



**Figure 5.4:** The data simulation approach is optimized by parallel utilization of the CPU cores on the system.

in the system. Once signals have been generated for all structures, the training data is saved in one large file in the *npy* format. Therefore, the *python-powder-diffraction* package not only provides the required functionality for the simulation of realistic powder patterns but also offers scripts to generate datasets rapidly on a large scale.

## 5.3 Neural Network Benchmark

### 5.3.1 Package Overview

The application of neural networks necessitates a robust and efficient framework that includes the implementation of the models' functionality, including convolutional layers. To this end, the most common libraries for deep learning are available in Python: *PyTorch* and *TensorFlow*. Accordingly, in the context of this thesis, neural networks have been implemented within the *TensorFlow* ecosystem, which provides the optimized implementations of pre-defined layers, loss functions, and various optimizers.

Nonetheless, exemplary data is necessary to train or evaluate the neural networks. Accordingly, a synthetic dataset to evaluate distinct network configurations has been introduced in Chapter 4, which has been implemented in the *spectra-network-benchmark* framework and is available in the following repository: `https://github.com/jschuetzke/spectra-network-benchmark`.

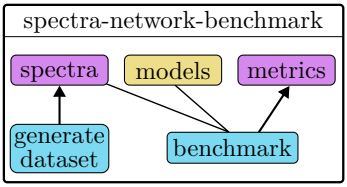



**Figure 5.5:** The structure of the *spectra-network-benchmark* framework. As the core component, synthetic datasets can be generated using the *generate dataset* script. Based on this dataset, different neural network models can be evaluated using the *benchmark* script, which provides the metrics to quantify the performance of the models.

Figure 5.5 provides an overview of this framework, which includes scripts and methods for simulating the patterns and several model implementations, with the user-executable scripts highlighted in blue. Currently, implementations of neural networks from previous studies, as well as the proposed architecture, are included in the *spectra-network-benchmark* framework. Thus, the framework is particularly useful for benchmarking the performance of novel architectures against established models.

### 5.3.2 The Synthetic Spectra Benchmark Framework

While the simulation of diffraction patterns or spectra necessitates the integration of a simulation tool, arbitrary patterns can be generated more rapidly, as the placement and height of peaks are not related to the physical properties of a structure. Therefore, it is possible to define a unique pattern unrelated to a structure or data evaluation domain based on the positions $pos$ and heights $his$ of its peaks. Here, $pos_i$ and $his_i$ describe the position and height of the $i$-th peak in the pattern with $n_{\text{peaks}}$ total peaks. Furthermore, defining the length $l$ of the signals is necessary. Subsequently, generating $n_{\text{signals}}$ varied patterns rapidly using Algorithm 2 is possible.

The discrete peak information is convolved with different Gaussian profiles ($G$) to generate signals with varying FWHMs. Similar to the XRD pattern simulation

**Algorithm 2** Generation of varied patterns for artificial spectra classes.

**Require:** $l$ ▷ signal length
**Require:** $n_{\text{signals}}$ ▷ number of samples to generate
**Require:** $pos,\ his$ ▷ positions and heights of the peaks
**Require:** $pos_{\text{shift}},\ his_{\text{shift}}$ ▷ acceptable shifts

```
for s = 1, 2, ..., n_signals do
    for i = 1, 2, ..., n_peaks do
        pos_i ← pos_i + Random(−pos_shift, pos_shift)
        his_i ← his_i + Random(−hi_shift, hi_shift)
    end for
    signal_s ← pos, his
    signal_s ← sample_s ∗ G                ▷ convolve kernel
end for
```

approach, the *SciPy* function *gaussian_filter1d* provides an optimized implementation of this convolution operation. Therefore, the properties of the generated peaks can be directly manipulated by explicitly setting a range of acceptable FWHM values that are used to produce the broad peak shapes.

Generating samples that act as negative examples is essential to train and evaluate neural network models. Thus, additional patterns can be generated through the simulation of different fingerprints, with peak positions and heights contradicting the properties of the pattern to identify. In this case, Algorithm 2 can be utilized without modifications by providing alternative peak positions $pos$ and heights $his$. Alternatively, it is possible to generate modifications of the original pattern by the addition of supplementary peaks.

The approach of generating a synthetic dataset that contains positive and negative examples is implemented in the *generate_dataset.py* script. Furthermore, this repository provides several implementations of network architectures in the *model_implementations* directory, including the models developed by Szymanski et al. [43] and Lee et al. [16]. Using the *benchmark.py* script, those models can be trained and evaluated on the synthetic signal samples.

### 5.3.3 Framework Utilization

Concurrently, the *spectra-network-benchmark* framework contains implementations for some of the established neural network architectures. In case, further neural networks are developed, those can be added to the framework. To do so, a new file is added to the *model_implementations* directory, and the model's name is inserted into the *benchmark.py* script. Then, the performance of the model and the count of the model's parameters is determined. The training progression and the metrics of the models are tracked on the weights & biases platform. On the `https://wandb.ai/jschuetzke/model_selection` project page, the logs of the different network training procedures are available.

It is also possible to evaluate the models for the detection of alternative fingerprints. Here, signals with a length of 5000 data points were used and the target fingerprint was defined with two peaks at positions 4290 and 4700 and relative heights of 74.4% and 100%. A variation of 100 data points (positions) and 10% peak height (absolute) was also included, while ensuring that all peaks were still detectable. Thus, one can leverage the *spectra-network-benchmark* framework to generate another data set with alternative values and benchmark the models against these synthetic signals.

## 5.4 Crystal Structure Identification Framework

### 5.4.1 Package Overview

The overarching objective of this thesis extends beyond the analysis of synthetic patterns using neural networks; it aims to integrate these networks into material discovery systems, thereby enabling the automatic identification of novel materials in diverse datasets. A comprehensive framework has been developed to pursue this goal, as presented in Chapter 2. This framework contains several methods tailored for several critical stages: the generation of training data, the design and training of neural network models, and the analysis of measured data. Accordingly,

the *crystal-id* framework provides the technical implementation of the presented methods and allows for straightforward integration of the framework into existing material discovery systems. The *crystal-id* framework is available in a public GitHub repository:
`https://github.com/jschuetzke/crystal-id`.

Figure 5.6 provides an overview of the *crystal-id* framework. As its core functionality, the framework enables the training of a neural network model to identify a target material using the user-executable scripts, which are highlighted in blue. To accomplish this, only a CIF description of the target material is required and the user potentially has to modify the parameters of a configuration file. After training, the model is available to rapidly analyze measured scans, providing the results in both human-readable and machine-readable formats. Accordingly, Section 5.4.2 explains the scripts of the framework in more detail. Then, Section 5.4.3 describes the procedure of integrating the framework into existing material discovery systems.

### 5.4.2 The Crystal Identification Framework

Generally, the framework can be separated into three fundamental scripts: the training data generation, the model training, and the analysis of measured data. Accordingly, the crystal-id framework contains three Python scripts that incorporate this functionality:

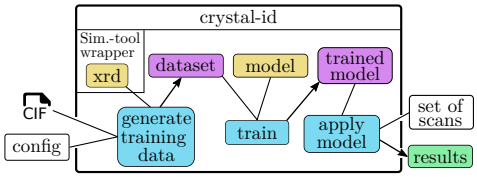



**Figure 5.6:** The structure of the *crystal-id* framework. The *generate training data* and *train* scripts produce a trained neural network model that can be used for the analysis of measured scans using the *apply model* function.

### generate training data

The training data generation script is used to create a comprehensive dataset of simulated scans that can be used to fit the parameters of the neural network model. Given that this thesis mainly focuses on analyzing XRD patterns from material discovery experiments, the methodologies developed for the *python-powder-diffraction* package are employed for this data generation. Hence, for the provided description of a material structure, the script primarily generates varied diffraction patterns that represent the variations of the characteristic fingerprint concerning peak positions, intensities, and shapes. However, the *python-powder-diffraction* library does not include functionality to represent the experimental outcomes that failed to produce the target material. To address this, the data generation script extends the functionality of the *Powder* class to include simulations of alternative structure configurations, multi-compound mixtures, and amorphous scans according to the methodology presented in Section 3.4.

To initiate the generation of training data, a description of the target structure in CIF format is required. In addition, it is crucial to specify the elemental properties of the signals to be simulated in the *config.yml* file, such as the measurement range, step size, and threshold for the acceptance of impurity phases alongside the target material. An exemplary *config.yml* file can be found in the Appendix. By default, the data generation script is programmed to produce a dataset comprising 1000 simulated signals. Out of these, 500 signals are positive examples displaying the fingerprint of the target material, while the remaining 500 constitute negative examples. The composition of these negative examples is carefully structured: 10% are amorphous patterns, 20% represent alternative structures, and the remaining 70% are a mix of the target fingerprint with impurities. This distribution of samples has been empirically found to yield the most effective results for the training process.

Additionally, the configuration file offers the flexibility to modify the default properties of the simulation. For instance, users can adjust the range of FWHM values to generate patterns with broader peaks. The extent of variation in lattice parameters and peak heights is also customizable, allowing for the generation of

fingerprints that are even more different from the CIF structure. Furthermore, the file includes parameters to define the level of noise and the ratio of background intensities in the simulated scans. However, it is important to note that the default values have been carefully defined based on the properties of measured scans from the RRUFF database, and in most cases, these defaults are sufficient and do not require alteration.

### train

The training of the developed neural network model, as described in Section 4.2.5, is enabled by the *train.py* script. This script leverages the *TensorFlow* library for the implementation of the network, building on optimized code instead of implementing the functionality of the deep learning model from scratch. A key feature of this model implementation is its ability to take the length of the signal as input, allowing for the initialization of a model with parameters tailored to the dataset at hand. Furthermore, the configuration of the model can be easily modified by providing alternative arguments during its initialization. For instance, adjustments such as adding more filters, increasing kernel sizes, or reducing the number of layers can be achieved by supplying different values to the "filters", "kernel", or "layers" arguments, respectively. However, it is noteworthy that the default parameters match the model configuration as described and are, therefore, optimized.

Initially, the training script imports training and validation datasets and applies a minimum-maximum scaling to the synthetic signals. This ensures that the input data is appropriately scaled for neural network processing, a standard practice in the training of models for analyzing XRD patterns or Raman spectra. To train the models, an Adam optimizer with a learning rate of $1 \cdot 10^{-4}$ and default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \cdot 10^{-7}$, is utilized, which is also provided by the *TensorFlow* library. During training, the training samples are used for adapting the model's parameters and the validation loss is observed to halt the training process once a plateau in loss minimization is observed. This strategy is crucial in preventing overfitting on the training data, thereby enhancing

the model's ability to generalize effectively to new, unseen samples. Finally, the trained model is saved, so it can later be loaded into separate data analysis scripts.

#### apply model

The final step of the crystal-id framework involves the application of the trained models to the measured data. This is facilitated by the *apply_model.py* script, which has been designed for the automatic categorization of a list of measured signals. To execute the data analysis procedure, the script requires two key inputs: the specification of the model and the directory containing the scans. Scans are typically saved in individual text files, e.g., in the "xyd" or "txt" format, with columns specifying the measurement steps and detected intensity counts. Upon running the script with these inputs, it systematically processes each scan, applying the trained neural network model to determine the categorization of each diffraction pattern. The output of this process is encapsulated into a single text file, *results.csv*, where each diffraction pattern is labeled either with a "1", indicating a match with the target structure, or with a "0" in cases where it does not match. This streamlined approach enables efficient and accurate classification of measured data, offering a valuable tool for researchers in the field of material discovery.

### 5.4.3 Framework Utilization

To effectively utilize the crystal-id framework, only few steps are necessary. The framework is readily accessible for download from Github under the following URL:
`https://github.com/jschuetzke/crystal-id`.

Once obtained, the next step is to install the required Python packages, such as *TensorFlow* and *NumPy*, which are listed in the *requirements.txt* file. The initial operational step involves selecting or formulating a CIF that accurately describes the target crystal structure. Depending on the experiments, this structure can either

be representative of a stable region within the compositional space being explored in experiments or is obtained from a database that provides computed material structures. Additionally, users need to adjust the parameters in the *config.yml* file to align with their specific measurement configuration. This preparatory process ensures that the framework is precisely configured to analyze the materials and experiments of interest, laying the foundation for accurate and effective analysis.

Once the preliminary setup is completed, the user can proceed with running the Python scripts that form the core of the crystal-id framework, as conceptualized in Figure 5.6. The process begins with the execution of the *generate_training_data.py* script. Here, the user specifies the name of the material system and the corresponding CIF file as input arguments when running the script. Next, the model training script *train.py* is executed, where the only requirement is to specify the name of the material system for which the exemplary signals have been generated in the first step. Finally, the application of the trained model is carried out using the application script *apply_model.py*. In this final step, the user inputs the name of the material system and the directory containing the measurement files. This sequential execution of scripts ensures a streamlined workflow from data generation and model training to the practical application of the model on real-world data.

The outcomes of the automated analysis conducted by the crystal-id framework are conveniently compiled in a *results.csv* file, enabling straightforward interpretation of the results. The provision of results in a machine-readable format opens up possibilities for integrating the model's predictions into an iterative experimental approach. This feature of the framework makes it particularly conducive to simple integration into existing material discovery systems, whether they operate on a semi-automatic or fully automatic basis. Such integration not only streamlines the experimental workflow but also enhances the efficiency of the material discovery process by leveraging the predictive capabilities of the model.

The crystal-id framework is also designed with future extensibility in mind, particularly for the integration of tools that simulate Raman spectra. Currently, the simulation of XRD scans is facilitated by the *XRDCalculator* class from

the *pymatgen* library, which requires inputs such as material structure, radiation wavelength, and measurement range to predict the corresponding diffraction peak positions and intensities. This functionality is implemented in the "wrapper" directory of the crystal-id repository, where the "xrd.py" simulation tool is currently the only available module. Thus, the framework's modular design allows for the potential addition of a Raman simulation module under "raman.py". Such a module would similarly take the material structure and measurement range inputs to produce a list of peak positions and intensities for Raman spectra. Thanks to this modular design, the crystal-id framework is ready for seamless expansion as soon as the requisite tools for Raman spectrum simulation become available.

# 6 Application

## 6.1 Overview

In the previous chapters, a comprehensive framework for the automated identification of novel materials through their characteristic fingerprints in experimental data has been presented. The application of this framework aims to address the bottleneck of manual data analysis that currently limits the efficiency of various high-throughput material discovery systems. Therefore, this chapter is dedicated to presenting the application of the novel framework on datasets from various material discovery experiments. Depending on the characterization technique and the measurement modalities employed to acquire the signals, the fingerprints of the target material can vary greatly. In addition, a wide variety of synthesis outcomes can occur in the experiments, resulting in a broad spectrum of distinctive patterns for analysis. Hence, the framework is applied to three distinct experimental datasets, and its performance, in terms of accuracy and time efficiency, is compared to that of a manual analysis.

Section 6.2 describes the analysis of high entropy oxide materials, with the primary objective of identifying compositions that crystallize into the fluorite structure. Following this, Section 6.3 focuses on the analysis of experimental data related to the formation of the disordered rocksalt phase. In this section, the framework is particularly instrumental in identifying scans that exhibit the presence of target materials amidst impurities, a necessity due to the inability to synthesize a pure sample in the study. Section 6.4 then presents the analysis of experimental data aimed at producing Yttrium Barium Copper Oxide, which is an established superconducting material with many use cases. Nonetheless, the experimental series

tests the synthesis process of this material with respect to precursors and synthesis temperature with the aim of increasing the yield of the target structure. Accordingly, most of the scans in this dataset do not contain the intended material or it is only present next to impurity phases. Finally, Section 6.5 critically evaluates the accuracy of the predictions made by the framework. It also presents a comparative analysis between the time required for automated analysis and traditional manual phase identification, highlighting the efficiency and effectiveness of the automated approach in handling complex material analysis.

## 6.2 Identification of Fluorite Structures

### 6.2.1 Description of Dataset

In the study conducted by Velasco et al. [53], a comprehensive investigation was carried out on a multi-compound oxide material system with end members lanthanum (La), cerium (Ce), yttrium (Y), praseodymium (Pr), and samarium (Sm), which form the edges of the multi-dimensionional composition space. This system is capable of forming a variety of phases, as documented in several studies [90, 91]. Notably, the end members $CeO_2$ and $PrO_2$ are known to form a fluorite structure, while $Sm_2O_3$, $Y_2O_3$, and $La_2O_3$ typically crystallize in the bixbyite structure [90], with these formations well-documented in the ICSD database. However, experimental analyses of many combinations of these members are still limited. Concurrently, the Materials Project (MP) database includes entries such as $CePrO_4$, identified as "mp-1226481". These entries represent computationally predicted structures, yet they have not been confirmed as stable and exhibit structural characteristics distinct from those previously reported.

Accordingly, the study aimed to determine the phases formed from different compositions of the end members in the material system, which have not been reported in previous studies. Correspondingly, Figure 6.1 presents a three-dimensional phase diagram featuring four of these end members, intentionally excluding $Y_2O_3$ to avoid the complexity of visualizing an additional dimension. The phase diagram

illustrates which compositions are reported to form a fluorite (green triangles) or bixbyite (blue pentagons) structure, as well as the additional compositions (gray circles) examined in the study. In total, Velasco et al. synthesized 106 samples with distinct compositions (pure end members plus mixtures).

To identify the crystalline phases that were formed by each composition, all synthesized samples were analyzed using an XRD instrument. Accordingly, analysis of the XRD scans allowed Velasco et al. to determine the phases for the distinct compositions. This dataset, which has been made available by the authors of the study, also presents an interesting case study for the application of the framework. The objective is to identify the structure that has formed during synthesis for the various compositions tested in the dataset. Through manual analysis of the XRD patterns, the baseline for human performance can be established in the context of this thesis. The novel framework is then applied to automatically analyze the scans and the predictions of the integrated neural network model are compared to the results of the manual analysis. While the original study determined the exact phase or mixture constitution for each composition, the framework is applied here to demonstrate the rapid identification of samples consistent with the fluorite structure.

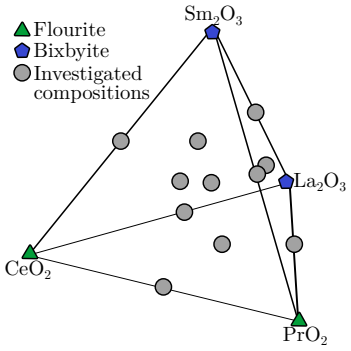


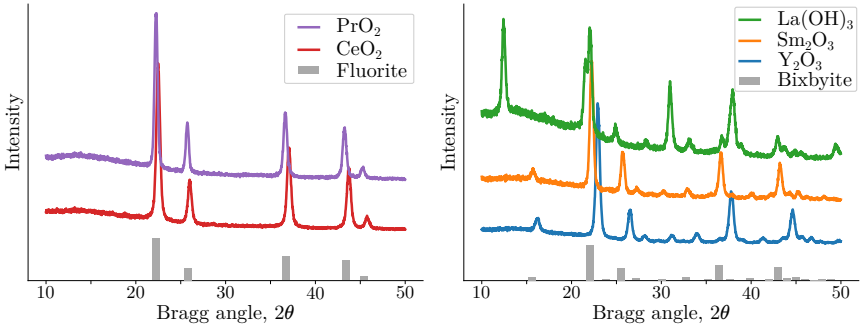

**Figure 6.1:** Quaternary phase diagram of the (Ce,Pr,Y,Sm,La)$O_2$ material system (excluding $Y_2O_3$). The gray circles indicate the compositions that were specifically examined in the study.

### 6.2.2 Manual Analysis

Figure 6.2 displays selected XRD patterns from the acquired dataset. Specifically, Figure 6.2a showcases the measured scans (red and purple) for the samples containing the pure $CeO_2$ and $PrO_2$ end members. These phases are reported to form a fluorite crystalline structure [90], so the characteristic fingerprint for this phase is simultaneously presented as gray bars. The samples in this study were scanned using Ga-jet radiation, featuring a K$\beta$ wavelength of 1.2079 Å, so the exact positions and heights of the diffraction peaks were simulated using the *pymatgen* library for ICSD entry 24887 and a matching wavelength. Accordingly, the highly symmetric structure is distinctively characterized by the presence of only five peaks at positions 24, 26, 37, 44, and 45 degrees $2\theta$.

The measured XRD patterns of the $CeO_2$ and $PrO_2$ samples align more or less precisely with the positions of the simulated diffraction peaks. However, there are minor deviations from the simulated pattern, most noticeable in the peaks occurring at higher angles. This is due to the differences in content and proportions of the unit cells of the distinct structures. For instance, the ICSD entries for $CeO_2$ and $PrO_2$ show unit cell edge lengths of 5.47 and 5.73 Å, respectively. Moreover, these elements exhibit slight variations in electron density, resulting in minimally different scattering factors. As a result, while the diffraction patterns capture the characteristic features of the fluorite structure, the peak positions and intensities can vary greatly depending on the composition, which complicates the analysis of the measured scans.

Furthermore, Figure 6.2b illustrates the measured scans (blue, orange, and green) for the samples containing the remaining end members of the five-dimensional composition space. In this study, $La(OH)_3$ was used instead of $La_2O_3$, but this has no meaningful effect on the formation of the intermediate compositions, such as $LaSmO_3$. $Sm_2O_3$ and $Y_2O_3$ are reported to crystallize in the bixbyite phase [90], so the fingerprint of this phase is additionally displayed as gray bars (simulated for ICSD entry 8493). The bixbyite structure represents a lower symmetry version of the fluorite structure, so in addition to the five characteristic peaks observed in Figure 6.2a, there are more unique distances of planes in the bixbyite structure that




**(a)** XRD patterns for the samples with compositions $CeO_2$ and $PrO_2$ in comparison to the simulated pattern for an exemplary fluorite structure (ICSD entry 24887).

**(b)** XRD patterns for the samples with compositions $Sm_2O_3$, $Y_2O_3$, and $La(OH)_3$ in comparison to the simulated pattern for an exemplary bixbyite structure (ICSD entry 8493).

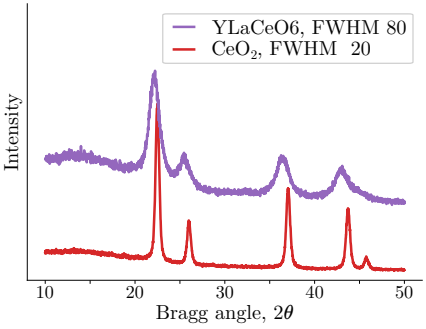
**Figure 6.2:** XRD patterns for the samples that contain the pure end members of the phases that constitute the multi-component composition space.

result in diffraction peaks. Although there are minor differences between the peak positions in the measured and simulated patterns, manual analysis of the XRD patterns attributed to the $Sm_2O_3$ and $Y_2O_3$ samples allows the conclusion that these samples formed the bixbyite structure. In contrast, the sample containing the $La(OH)_3$ material has a diffraction pattern that is significantly different from the bixbyite fingerprint, so it can be concluded that this sample is not consistent with either the bixbyite or the fluorite structure.

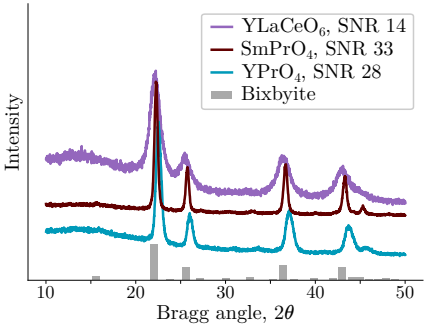
To accurately determine the crystalline phases present in each sample, the XRD dataset underwent a thorough manual examination using the QualX software [39] and entries from the ICSD. Accordingly, the exact structure of the samples was successfully identified. For example, the sample containing the end member $La(OH)_3$ is identified as having a fingerprint similar to an apatite structure, with a hexagonal crystal system and space group $P6_3/m$. In total, 88 of the 106 samples were conclusively determined to contain the fluorite structure without impurities, while others showed a fingerprint similar to the bixbyite structure or even patterns that allowed the sample to be identified as multiphase. This manual analysis of the 106 samples was completed in approximately 90 minutes.

In addition to the determination of the underlying phases for each sample, the manual analysis revealed that there is a substantial variation in the properties of the scans. Figure 6.3 shows XRD patterns with varying peak widths or noise levels that highlight this variety of properties in the dataset. For instance, the scan of the pure $CeO_2$ sample (depicted in red) is characterized by peaks with an FWHM of around 20. In contrast, the $Y_{0.33}La_{0.33}Ce_{0.33}O_2$ sample, depicted in purple in Figure 6.3a, displays broad diffraction peaks with an FWHM of 80.

Additionally, the high level of noise in some signals (low SNR) complicates the manual analysis of the scans. In some cases where a bixbyite structure was formed, the diffraction peaks necessary for drawing this conclusion are barely recognizable, as displayed in Figure 6.3b. This is exemplified in the XRD scans of $Sm_{0.5}Pr_{0.5}O_2$ and $Y_{0.5}Pr_{0.5}O_2$ compositions, shown in cyan and dark red, where subtle intensity elevations between 30 to 35 degrees hint at additional diffraction peaks. Accordingly, the manual analysis concluded that those compositions form a bixbyite structure. However, in scans with a lower SNR, such as that of $Y_{0.33}La_{0.33}Ce_{0.33}O_2$, these minor diffraction peaks cannot be conclusively identified due to the high level of noise, indicating that the measurement configuration in some cases does not yield sufficiently high-quality data for clear analysis. In




**(a)** XRD patterns with dissimilar FWHM characteristics.




**(b)** XRD patterns for intermediary compositions of the phase diagram where few additional peaks are barely noticeable.

**Figure 6.3:** Selected XRD patterns from the multi-component dataset that highlight the variety of signal properties.

this particular case, no additional diffraction peak indicative of the bixbyite structure could be detected above the noise level, so the $Y_{0.33}La_{0.33}Ce_{0.33}O_2$ sample was identified as consistent with the fluorite structure.

Notably, Velasco et al. also reported the crystal structure determined from the analysis of the XRD patterns in their study. The results of the manual analysis performed in the context of this thesis are mostly consistent with the results reported by the authors of the study. However, deviations are observed for a small subset of samples (< 5), where patterns are categorized as bixbyite instead of fluorite, or vice versa. These discrepancies primarily arise due to low SNRs in the respective signals, where the elevated noise levels potentially obscure additional diffraction peaks. Thus, the results of the manual analysis conducted in this thesis are assumed to be correct in the following sections.

### 6.2.3 Application of the Novel Framework

As a comparison to the manual analysis, the acquired XRD patterns are also analyzed using the novel framework. Due to the universal design of the framework, it can be applied for the scans of the different compositions that explore the high-dimensional composition space without requiring modifications. The automated approach is designed to identify those samples with a fingerprint that matches a target phase. For this specific dataset, the fluorite structure has been identified as the target phase in the context of this thesis. Thus, the goal is to identify the same 88 samples that were previously classified as fluorite through manual analysis while categorizing the remaining samples as not conforming to fluorite.

The application of the framework requires two essential components: a CIF of the target structure and a definition of whether the presence of impurities is acceptable. In this specific case, the process was straightforward: a CIF file was sourced from the ICSD entry that represents the fluorite structure (code 24887). Given that the presence of additional diffraction peaks could indicate a structure differing from the target phase, simulated patterns were generated that aligned precisely with the characteristic fingerprint of the fluorite structure. Any deviation from this pattern

indicates the presence of impurities or alternative structures, e.g., bixbyite, so the dataset has to represent this information accurately to enable the training of a robust neural network model.

Generally, the framework allows for generating synthetic diffraction patterns, which have identical measurement modalities (scanning range, step width) as the measured scans. This approach ensures that the neural network, once trained, can be directly applied to the measured data. The measurement range for the acquired scans spanned from 10 to 50 degrees $2\theta$ with a step width of $0.015° \Delta 2\theta$. However, the presence of broad diffraction peaks is a significant challenge in this particular dataset. The neural network was designed with signal characteristics in mind that match those found in most measured signals, which includes peaks with FWHMs in the range between 5 and 60. Yet, the dataset contains signals with peak shapes that are even broader, as shown in Figure 6.3. To effectively address this, a decision was made to theoretically increase the step size for the models to $0.03° \Delta 2\theta$, which effectively halves the width of all peaks.

Accordingly, training data was simulated in the range from 10 to 50 degrees $2\theta$ for the original Ga-jet radiation wavelength, conserving the original measurement range, with a step width of $0.03° \Delta 2\theta$, resulting in 1334 data points. This adjustment allowed the neural network model to handle the broad diffraction peaks characteristic of this data better. Additionally, it was crucial to simulate scans with a high level of artificial noise to depict those scans with a low SNR accurately. This step was essential to ensure that the model was well-trained to recognize patterns even in data where the signal quality was impaired, thereby enhancing the robustness and applicability of the model under varying data quality conditions.

The entire process of simulating training data and training the model was efficiently completed in approximately five minutes. Analysis of the 106 samples was performed within less than a second by the trained model. Upon comparison with manual analysis techniques, it was found that the model's predictions aligned with the manual evaluations for 104 out of the 106 samples, demonstrating a remarkable accuracy of approximately 98%.

Notably, scans with low-intensity, additional diffraction peaks, and high SNR, similar to the grey and pink patterns shown in Figure 6.2b, led to misclassifications by the model. This discrepancy may highlight a limitation in the optimization approach: the model must balance between compensating for high noise levels in some scans and detecting minor diffraction peaks of similar magnitude in others. However, the challenge presented by low SNR scans also impacted the manual evaluation, suggesting that the selection of acquisition parameters should be carefully optimized to ensure high SNR in all measured scans. Such an approach would allow for more definitive conclusions to be drawn from the analysis, enhancing both manual and automated evaluations' reliability and accuracy.

Based on the classification of the samples with distinct compositions, the neural network model has simplified the creation of an insightful phase diagram, showcased in Figure 6.4. This diagram specifically illustrates the ternary phase system involving $CeO_2$, $PrO_2$, and $La_2O_3$. Here, the compositions that have been synthesized as part of the original study are highlighted by the symbols in the diagram that also signify the model's predictions. Notably, the analysis of the samples reveals that the sample with composition $Ce_{0.5}Pr_{0.5}O_2$ exhibits the fluorite structure. Similarly, it was found that a combination of all three end members ($Ce_{0.33}Pr_{0.33}La_{0.33}O_2$) also resulted in the formation of the fluorite structure. However, compositions with 50% $La_2O_3$ displayed XRD patterns that deviated from the typical fluorite structure pattern.

The phase diagram was manually compiled using the model's predictions for the XRD patterns of the samples with distinct compositions. Consequently, the model only evaluated these specific positions within the diagram. To complete the phase diagram, there are two approaches: either the phase boundaries can be approximated based on these points, or additional samples are generated and analyzed to fill in the gaps in the diagram. Upon obtaining XRD patterns for these additional samples, the same neural network model can be applied without the need for retraining, rapidly providing the corresponding structures for the distinct samples. This feature underscores the utility of the model in experiments aimed at accurately determining phase boundaries in complex material systems.

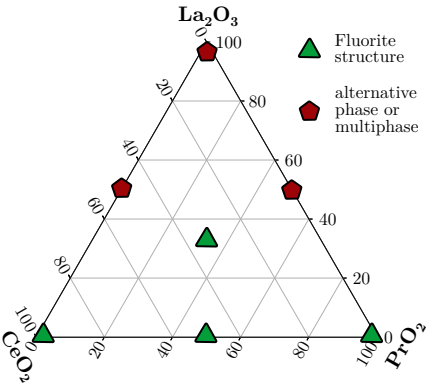



**Figure 6.4:** Partial phase diagram of the multi-component material system, as generated by the classification of the neural network model. The green triangles indicate the formation of the fluorite phase, while the red pentagons delineate areas where the fluorite phase is not formed exclusively.

## 6.3 Identification of Disordered Rocksalt Structures

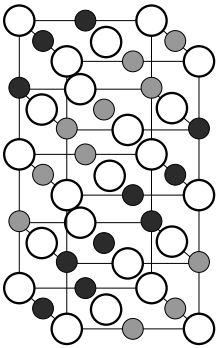
### 6.3.1 Description of Dataset

In recent years, the interest in disordered rock salt (DRX) structures has surged, particularly due to their inherent cationic disorder and distinct electrochemical properties, making them highly sought-after for battery applications [92]. DRX structures emerge as potent cathodes with high energy densities that have the potential to elevate contemporary battery technology. These materials are typically represented as $LiMO_2$, where "M" denotes various transition metal species, such as iron, nickel, or titanium. As the nomenclature suggests, these compounds adopt a rock salt-like configuration, as introduced in Figure 1.8. The term disorder refers to the intermixing of lithium and transition metal in certain positions within the lattice [92].

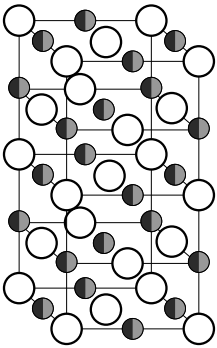
Figure 6.5 presents the lattices of two distinct rock salt structures. In this representation, black or gray circles symbolize lithium and transition metal sites, whereas

white circles correspond to oxygen positions within the lattice. The structure in Figure 6.5a displays the ordered form, characterized by alternating layers of transition metals and lithium. Conversely, Figure 6.5b illustrates the disordered variant where lithium and transition metal sites are interspersed uniformly throughout the lattice. Circles featuring mixed shades indicate sites that can be occupied either by lithium or a transition metal atom, with occupation probabilities aligning with the specific DRX composition ratio [93].

Transition metals, a diverse group of elements predominantly located in the d-block of the periodic table, offer a vast array of candidates that can be incorporated into the DRX structure. Each of these metals possesses unique electrochemical properties that can influence the overall performance of the DRX structure when used as a cathode in batteries. Given the abundance and variety of transition metals, combined with the possibility of mixing multiple metals in varying ratios, the potential compositional combinations within the DRX framework become vast. Consequently, plenty of compositions need to be synthesized and thoroughly



**(a)** Layered structure with a regular order.



**(b)** Disordered rocksalt (DRX) structure with equivalent sites.

**Figure 6.5:** Distinction between ordered and disordered rocksalt structures that are commonly used as battery materials. The white circles represent oxygen sites, and the black and gray circles depict lithium and transition metal sites. Adapted from [93].

evaluated to comprehensively understand their electrochemical properties, ensuring the optimal selection and design of cathode materials for advanced battery applications [92].

A noteworthy avenue of research has been the fluorination of DRX structures. By introducing fluoride ions in place of some oxygen sites, one can further enhance the electrochemical properties of these materials. This fluorination not only modifies the lattice energetics but also impacts the electronic and ionic conductivity of the structure, leading to potentially improved battery performance [94]. However, only small amounts of fluorite substitutions can be achieved using conventional synthesis routes [95].

To address this limitation, Szymanski et al. [61] conducted a comprehensive study to explore different precursor sets and experimental conditions that are beneficial to the formation of the highly fluorinated DRX phase. In particular, the $Li_{1.2}Mn_{0.4}Ti_{0.4}O_{1.6}F_{0.4}$ phase has been identified based, which was pinpointed based on predictions from a simulation tool. Most notably, the study involved the use of an *in situ* instrument to analyze the formation of the DRX phase while performing the synthesis. These advanced tools enable the continuous acquisition of powder diffraction patterns while simultaneously monitoring the experimental conditions, such as temperature. While such advanced instrumentation enables in-depth analysis of the phase formation processes during synthesis, this approach also results in the generation of large datasets containing dozens of patterns.

Accordingly, the primary objective of the data analysis procedure is to accurately identify those diffraction patterns that exhibit the characteristic features of the DRX phase. In total, 272 powder diffraction patterns have been acquired in the context of the *in situ* analysis. Therefore, this large-scale dataset, which has been provided by the authors of the study, offers another insightful case study for applying the novel material identification framework.

Nevertheless, Szymanski et al. noted that their experiments did not yield the successful synthesis of the target material without the presence of impurities. Consequently, the dataset proves beneficial in illustrating an additional feature of the novel framework: its adaptability. The automatic data analysis approach is

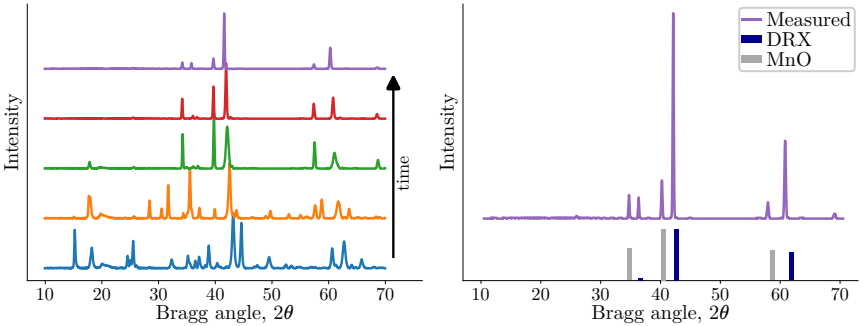
also proficient in recognizing the target structure even when additional materials are present. Accordingly, the following sections present the manual analysis of the dataset as well as the application of the automated framework, featuring a systematic comparison between the results obtained from both approaches.

### 6.3.2 Manual Analysis

Throughout the synthesis process, the XRD patterns measured show substantial variations, as depicted in Figure 6.6a. The patterns captured at the start of the synthesis predominantly reflect the precursor materials, including $LiMnO_2$, $Li_2TiO_3$, LiF, $MnF_2$, and C [61]. This is evident in the blue XRD pattern in Figure 6.6a, which displays a multitude of diffraction peaks attributable to the various phases present in the mixture. As the temperature increases during the synthesis, these precursors undergo reactions, leading to the formation of new phases. This transition is reflected by the reduction in the number of peaks in the green plot as compared to the initial XRD scan (in blue). As the synthesis progresses, the number of peaks continues to diminish.

Ultimately, the authors concluded that only the DRX phase is prominent in the synthesized sample, alongside the MnO impurity phase. Figure 6.6b illustrates the final XRD pattern obtained in the study, alongside the simulated positions and heights of the diffraction peaks for the MnO and DRX phases (depicted in blue and gray, respectively). While both phases share a similar rock salt-like lattice structure, they differ in their lattice parameters. The larger unit cell of the DRX phase causes the corresponding diffraction peaks to appear at higher angles $2\theta$ in the pattern. As a result, the peaks at 36, 42, and 61 degrees $2\theta$ can be attributed to the DRX phase, while those at 35, 40, and 58 degrees are indicative of the MnO phase.

While the MnO phase is well documented and represented in crystallographic databases, the DRX phase has only been identified through simulation tools. Consequently, conventional methods for analyzing the acquired patterns are not applicable without further information to characterize the DRX sample. To




**(a)** Selected DRX pattern from the dataset. The arrow indicates the progression of the patterns with respect to the acquisition time.

**(b)** Comparison of measured diffraction pattern with theoretical peak positions and intensities of the MnO and DRX phases. The simulated patterns were simulated using pymatgen [76] based on an MP entry of MnO (mp-19006) and a cif of the DRX structure provided by Szymanski et al.

**Figure 6.6:** Progression of the XRD patterns and temperature in the analyzed DRX dataset.

showcase the distinctive XRD pattern of this unique material, the description of the structure has been obtained from the authors of the study which enabled the simulation of the corresponding diffraction pattern.

Szymanski et al. included a detailed analysis of the dataset in their study, particularly focusing on the weight percentages of the various phases present in the samples Notably, the authors did not explicitly specify the time required for the manual analysis. However, despite providing detailed phase information, they did not establish a clear criterion for defining a "successful" synthesis outcome. In all measured XRD patterns, the DRX phase was consistently identified alongside impurities, leaving the term "successful" open to interpretation. Nonetheless, the dataset is considered here as a means to demonstrate the automated analysis performed by the novel framework, which provides a binary classification.

Thus, an arbitrary threshold has to be introduced in the context of this thesis to classify the dataset into "failed" and "successful" results. A straightforward approach to accomplish this is to consider the weight percentages reported by the authors. Consequently, the novel framework can be applied to the dataset, instructed to identify "successful" synthesis outcomes based on the identical definition. This

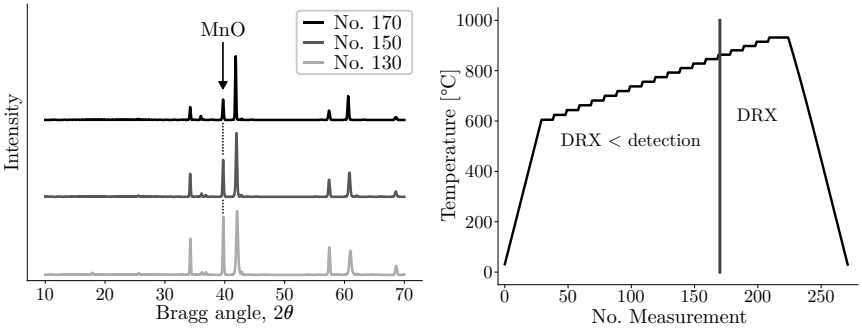
approach enables a direct comparison between the model's predictions and the results derived from manual analysis.

Figure 6.7a displays three scans from the dataset corresponding to measurement numbers 130, 150, and 170. In the case of scan number 130, the MnO impurity phase shows a peak at angel 40° $2\theta$ of nearly the same height as the highest peak of the DRX phase. As a result, the manual analysis determined that the MnO and DRX phases are present with approximately the same fractions ($wp_{\text{MnO}} \simeq wp_{\text{DRX}}$). In subsequent measurements, the intensity values of peaks linked to the MnO phase decrease, signaling a heightened prominence of the DRX phase. Consequently, it has been determined that the weight percentage of the MnO phase in measurement 150 is approximately two-thirds that of the DRX phase ($wp_{\text{MnO}} \simeq 0.65 \cdot wp_{\text{DRX}}$). This trend persists, with the MnO phase maintaining a weight percentage of 40% in measurement 170 ($wp_{\text{MnO}} \simeq 0.4 \cdot wp_{\text{DRX}}$).

As the primary objective of the study involves the synthesis of the DRX phase, a "successful" outcome should be characterized by instances where additional phases represent a minor fraction of the weight percentage compared to the primary phase. Therefore, the decision was made to establish an arbitrary threshold of ($wp_{\text{MnO}} \leq 0.4 \cdot wp_{\text{DRX}}$) to define the "successful" synthesis result. As reported by Szymanski et al., the specified criterion is satisfied for all scans beginning from number 170. For all scans later than number 170, the synthesis temperature was further increased, as shown in Figure 6.7b, resulting in diminishing weight percentages of the MnO phase [61]. Accordingly, the automated analysis should identify all scans numbered 170 or higher as aligning with the target material (plus acceptable impurities), while the preceding scans should be classified as "failed" synthesis outcomes, which is highlighted by the line in Figure 6.7b.

### 6.3.3 Application of the Novel Framework

Once the baseline for manual analysis of the dataset has been established (threshold 0.4), the novel framework is applied for automated analysis of the same XRD patterns. As described earlier, two essential components are required to apply the




**(a)** As the synthesis progresses, the DRX phase becomes more prominent. As an arbitrary threshold for classifying the patterns in the data set, the measurement with the number 170 was chosen as the cut-off point for identifying the DRX phase.

**(b)** Temperature progression during the synthesis. The previously defined cut-off at measurement number 170 is highlighted by the gray line.

**Figure 6.7:** Identification of patterns that contain DRX as the most prominent phase.

novel data analysis framework: a CIF of the target structure and a definition of whether the presence of impurities is acceptable. For the present dataset, the target material has been identified through computational tools and is not available from crystallographic databases. In this case, a CIF can typically be acquired from repositories like the Open Quantum Materials Database (OQMD) or the MP database. Here, the CIF has been identified from the original authors of the study who performed the simulations using custom *ab initio* software [61].

The XRD signals in the dataset show the diffraction patterns in the range between 10 and 70 degrees $2\theta$ with a step size of about $0.0133^\circ \Delta 2\theta$. Consequently, artificial XRD patterns are generated for the identical $2\theta$ range while considering the CuK$\alpha_1$ wavelength. However, the step size for simulations was intentionally increased to $0.02^\circ \Delta 2\theta$ to avoid the need to simulate intensities for rounded measurement steps. Furthermore, the described factor for acceptable impurity phase peaks is considered during the training data generation procedure. Apart from this, no further considerations are necessary for simulating diffraction patterns, as the scans exhibit high SNRs and narrow FWHMs.

Applying the novel framework took about five minutes, which involved the training data generation methods and model training. The analysis of the measured XRD

patterns was performed automatically by the trained neural network in less than one second. Furthermore, the predictions made by the model were consistent with the trends identified through manual analysis. Initially, the DRX fingerprint was not recognizable in the scans, resulting in predicted probability estimates equal to 0. However, as the synthesis progressed, the DRX phase began to form, and once a certain threshold was surpassed, the DRX phase was consistently predicted as present by the model. Notably, the model pinpointed measurement number 164 as the first "successful" synthesis result (according to the definition introduced in the context of this thesis). As a result, six out of the 272 samples were misclassified, leading to an accuracy score of 97.8%

Figure 6.8 effectively showcases two distinct XRD patterns that represent the DRX detection threshold as analyzed manually and as predicted by the model. On the left side, the first XRD pattern identified by the model as corresponding to the target structure is displayed. Conversely, on the right, Figure 6.8 presents the first XRD pattern that was manually identified as matching the target structure. The line drawn between corresponding impurity peaks in both plots underscores the slight difference between these two XRD scans. Similarly, the manually analyzed weight percentages are virtually indistinguishable ($wp_{\text{MnO,164}} \approx wp_{\text{MnO,170}}$).

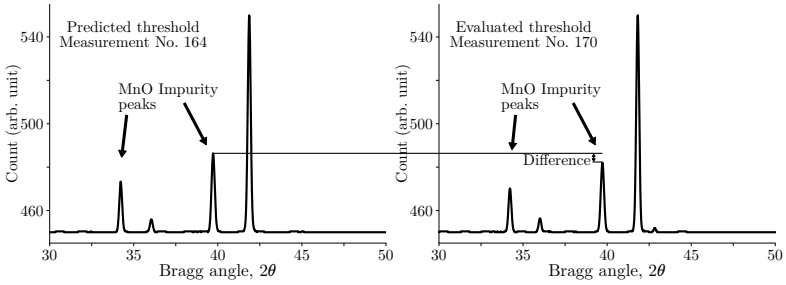



**Figure 6.8:** Comparison of DRX Phase Detection in XRD Patterns. Left shows the first XRD pattern predicted by the model to match the target structure, while the right plot displays the first pattern manually identified as such. The line indicates the minor impurity peak discrepancy between the two patterns, illustrating the model's sensitivity in comparison to manual analysis.

While there were six samples that the model misclassified, the disparity between these two specific scans is relatively negligible. This observation suggests that the trained model is somewhat less sensitive to the presence of impurities than the manual analysis. However, this discrepancy does not critically undermine the efficacy of the framework; it represents a minor inaccuracy that could be addressed in future refinements.

## 6.4 Identification of Yttrium Barium Copper Oxides

### 6.4.1 Description of Dataset

Yttrium Barium Copper Oxide (YBCO) is recognized as a superconducting material with notably high transition temperatures exceeding 77 K. Before the discovery of high-temperature superconductors, research in this field was largely limited to specialized laboratories equipped to handle experiments at lower temperatures. Consequently, YBCO has gained popularity as a superconductor, expanding the range of research and application possibilities significantly. The emergence of high-temperature superconductors like YBCO has facilitated the development of diverse applications, notably in electric motors, bearings, flywheels, and persistent current switches, utilizing the unique properties of thin film superconductors [96].

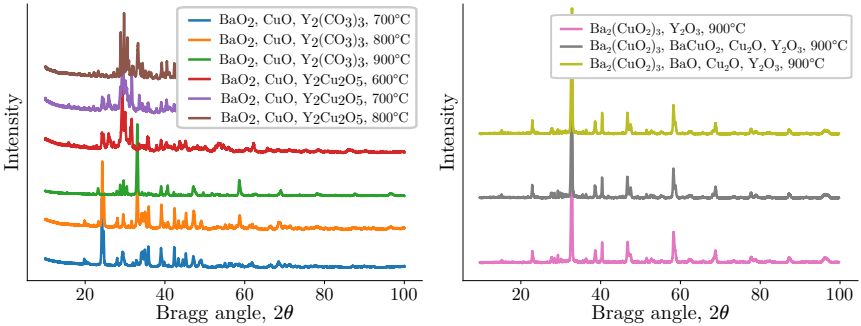
The synthesis of pure YBCO is not exceptionally challenging, but it allows for various precursor combinations and methods, all necessitating either high temperatures or prolonged experiment times [96]. To explore this, Szymanski et al. [97] performed a study to evaluate the formation of YBCO using diverse precursor combinations and experimental conditions. In particular, the composition $YBa_2Cu_3O_{6.5}$ was targeted. In this extensive study, samples were generated and analyzed using XRD instruments for 146 distinct configurations. Although most samples contained multi-compound mixtures, a few instances of successful synthesis of pure YBCO samples were recorded.

This dataset serves as the concluding case study for the newly developed framework dedicated to novel material identification. The previous sections demonstrated the application of the framework to explore a high-dimensional phase diagram or to identify the target phase in samples that also contained impurities. In contrast, this dataset allows the investigation of the framework's ability to identify the target phase, particularly in scenarios where it is rarely formed and in the distinct absence of impurities. Therefore, the framework is utilized in the following sections for the automated identification of samples containing the YBCO target phase.

### 6.4.2 Manual Analysis

In their original study, Szymanski et al. provided a detailed description of the phases present in their samples, including their respective weight percentages. However, the duration required for the manual identification of these phases or full profile matching was not specified in their report. A total of 146 XRD patterns were acquired, each recorded with a step size of $0.01° \Delta 2\theta$ in the $2\theta$ range of 10 to 90 degrees. These patterns were obtained using an XRD instrument equipped with a copper anode, which resulted in the presence of both K$\alpha_1$ and K$\alpha_2$ peaks. Notably, the measured signals exhibit a high SNR and standard peak shapes, with the FWHM of the peaks being 35 or lower.

Figure 6.9a displays representative XRD patterns, each annotated with labels denoting their respective precursor sets and synthesis temperatures. Within these exemplary patterns, the target phase is at least partially present in the signals colored orange, green, and brown, as reported by the original study. In contrast, the blue, red, and purple patterns exclusively exhibit alternative phases, highlighting the variability in phase formation under different synthesis conditions. Hence, Figure 6.9b illustrates three XRD patterns from samples that match the target material's structure, with each sample produced using distinct sets of precursors and synthesized at high temperatures, specifically at 900° C. Out of the array of samples studied, eight were identified as phase-pure compounds of the yttrium




**(a)** Exemplary XRD patterns of samples that either do not contain the target phase or it is only partially present.

**(b)** Selected XRD patterns for samples that have been identified as pure YBCO.

**Figure 6.9:** Selected XRD patterns from the dataset exploring the formation of the YBCO phase.

barium copper oxide phase. Consequently, the objective is to accurately identify these same eight patterns from the measured XRD scans by utilizing the novel material identification framework.

### 6.4.3 Application of the Novel Framework

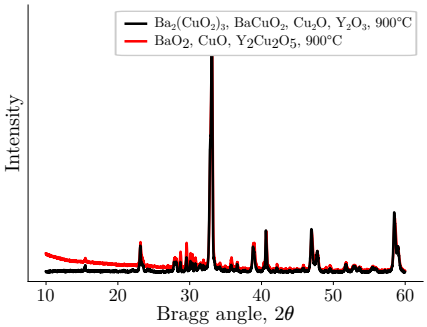
Concludingly, the application of the proposed concept is exemplified using the CIF description of $YBa_2Cu_3O_{6.5}$, which has been obtained from the ICSD (code 44117). The primary goal is the identification of samples devoid of impurities, hence the training data for the framework is simulated to reflect the formation of the target phase. As shown in Figure 6.9a, the relevant diffraction peaks predominantly span the range of 10 to 70 degrees $2\theta$. Accordingly, the simulations of the scans are restricted to this range, with a matching step size of $0.01° \Delta 2\theta$. Furthermore, no adjustments to the configuration are deemed necessary, since the levels of noise and the FWHM of the peaks align with the default settings of the framework.

The complete workflow enclosing data generation, model training, and data analysis was executed in less than five minutes. The trained model demonstrated proficiency in disregarding samples whose XRD patterns distinctly deviated from

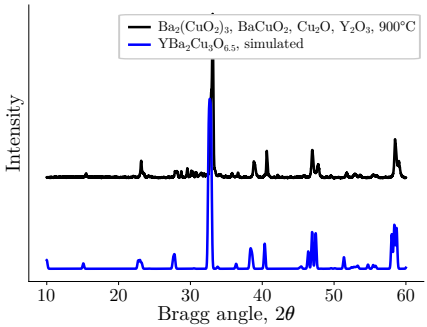
the characteristic fingerprint of the target material. Overall, the model correctly classified 143 of the 146 samples evaluated, achieving an accuracy of 97.9%. This demonstrates the model's high level of accuracy in identifying the target material through the characteristic fingerprint.

However, it recognized only five out of the eight samples successfully that were previously identified as phase-pure. Therefore, the misclassifications of the model are analyzed in more detail. Figure 6.10a displays two such patterns from samples that have been reported as phase-pure. Still, the model identified only the sample derived from precursors $Ba_2(CuO_2)_3$, $BaCuO_2$, $Cu_2O$, $Y_2O_3$ (black) as the YBCO material in its phase-pure form, while the other pattern (red) was not categorized as YBCO. Notably, the sample generated from the precursors $BaO_2$, CuO, and $Y_2Cu_2O_5$, depicted in red, shows markedly higher peak intensities at angles around 30° $2\theta$ ($\pm$ 2°). This observation underscores a substantial variation in the measured patterns, even among those samples that correspond to the YBCO phase according to the results reported in the original study.

Additionally, Figure 6.10b provides a visual comparison of the simulated diffraction pattern of $YBa_2Cu_3O_{6.5}$ (represented in blue) with an actual measurement




**(a)** XRD patterns of two samples identified as phase-pure YBCO. These signals highlight the notable variation in peak intensities around 30 degrees $2\theta$, even among samples with the same structure.




**(b)** Comparison of the simulated diffraction pattern of the YBCO phase (blue) with an actual measurement (black), emphasizing the absence of reflections around 30 degrees in the simulated pattern. This is indicative of minor impurities in the measured sample.

**Figure 6.10:** Comparison of XRD patterns that show the characteristic fingerprint of the YBCO phase.

from the study (shown in black). A key observation is that the simulated pattern shows no meaningful reflections around angle 30° $2\theta$. In contrast, the XRD scan of the measured sample largely aligns with the simulated fingerprint, except for the peaks that have been detected within this specific range. Consequently, it can be deduced that the presence of these peaks in the measured sample indicates the existence of minor impurities in the produced samples.

Thus, the predictions of the novel material identification framework are not necessarily wrong. The automated analysis attributed only those samples to the target material where the additional diffraction peaks in the region around 30° $2\theta$ remained minimal (five out of eight samples). On the contrary, the neural network model identified irregularities in the remaining three XRD patterns originally identified as YBCO, consequently categorizing those as "failed" experimental outcomes. However, these impurities are relatively minor and might not seriously impede the practical application of the superconducting material. Hence, while the model's predictions are visually confirmed as technically correct, the manual analysis failed to identify the same irregularities concerning the additional peaks in the XRD pattern.

## 6.5 Discussion of Results

The novel material identification framework was applied to three distinct datasets, demonstrating the simplicity and effectiveness of the developed methods. In each instance, the approach led to the rapid creation of a neural network model, capable of analyzing measured XRD scans within approximately five minutes for each dataset. Crucially, generating training data only requires a description of the target material in CIF format, complemented by a defined acceptance criterion for impurities in the samples. Notably, these prerequisites are attainable without the need for prior analysis of the measured scans. Furthermore, the successful implementation of the presented framework did not necessitate any specialized knowledge in the domain of deep learning. As a result, the application of this innovative framework does not demand specialized expertise and can be

seamlessly incorporated into the analysis process for various materials, enhancing its accessibility and usability in diverse research contexts.

The three datasets evaluated encompass a variety of radiation wavelengths, measurement ranges, and step sizes. Despite these differences, the application of the novel concept was demonstrated without necessitating significant modifications to the framework. This underscores the framework's versatility, as it proves adaptable across various sample characterization domains. Moreover, the approach consistently achieved accuracy scores of approximately 98%, showcasing the high level of performance and reliability of the automated analysis. This robustness further emphasizes the framework's potential for broad applicability in material identification and analysis.

Further in-depth analysis of the scans and instances of misclassification revealed that some of these errors can possibly be attributed to incorrect labeling during the manual evaluation process. Specifically, in the case of the fluorite structure, the diverse range of signal-to-noise ratios in the scans added complexity to the analysis, as potential additional diffraction peaks could be obscured by noise. In the DRX analysis study, an arbitrary threshold was initially set, and the automated analysis method identified a slightly different yet similar threshold, which is not necessarily incorrect. Ultimately, in the analysis of the YBCO dataset, the model demonstrated a heightened sensitivity to impurity peaks compared to manual analysis, further evidencing its precision when analyzing the measured XRD patterns.

Accordingly, the accuracy metric, while informative, does not fully capture the robust performance exhibited by the model. A critical aspect of its efficacy is that the model consistently avoided misidentifying multi-compound samples as the target phase, despite the varied appearances of their diffraction patterns. This highlights the effectiveness of the data generation approach employed, which, in addition to the characteristic fingerprints of the target phase, generates alternative signals for training the model. These alternative examples are efficiently created by introducing peaks at random positions within the signal, a process considerably faster than simulating and mixing XRD patterns of various phases present in the

evaluated material system. Therefore, the robustness of the automated data analysis is especially remarkable, as the network model was never exposed to the exact patterns of alternative structures or multi-phase compounds during its training phase. Yet, the presented approach consistently demonstrates high performance across diverse XRD scans, emphasizing the versatility and effectiveness of the presented concept for generating training data.

Furthermore, the comparison of automated analysis should not only take the achieved accuracy metrics into account but also the time efficiency of applying the automated analysis approach, offering a comprehensive evaluation of its overall effectiveness. Previous studies have emphasized the substantial manual effort required to identify materials by their fingerprint [44, 98, 99, 100], yet they fall short of quantifying the actual time needed for such manual analysis. For instance, Oviedo et al. [44] note that the process of acquiring measured data for a sample typically consumes about an hour, followed by an additional one to two hours for a comprehensive profile analysis of the XRD data, assuming that the phases present are already identified. However, it is important to recognize that full profile matching, a technique specific to XRD data, demands specialized expertise and is not universally applicable. Consequently, it is reasonable to infer that the task of phase identification, when relying on correlation coefficients or the figure-of-merit as metrics, requires less time.

For the dataset examining the formation of the fluorite structure within the multi-compound material system, the manual analysis of 106 measured scans was completed in approximately 90 minutes. Hence, times for the manual analysis of the other datasets can be approximated based using the baseline of 50 seconds per scan. However, it is crucial to note that this is a somewhat optimistic estimate, given that the diffraction pattern of the fluorite structure is particularly easy to discriminate. Therefore, it can be assumed that the identification of target materials with more complex fingerprints could be more time-consuming and challenging. Furthermore, this estimation assumes that an expert can maintain this pace without misinterpreting any signals, but it does not account for the potential fatigue from analyzing hundreds of scans. Such fatigue could severely affect the accuracy of the manual process [101]. Thus, while manual analysis is

**Table 6.1:** Comparison of the times required for manual or automated analysis of the different datasets evaluated in this thesis.

| Target phase | Fluorite | DRX | YBCO |
|---|---|---|---|
| Number of samples in dataset | 106 | 272 | 146 |
| Time required for manual analysis [min] | 90 | 227 (estimated) | 122 (estimated) |
| Time to apply novel framework [min] | 5 | 5 | 5 |

assumed to achieve perfect phase identification results for measured scans, analyzing large-scale datasets in automated material system platforms risks introducing errors due to human factors.

Table 6.1 showcases a comparative analysis of manual analysis versus the automated analysis framework in terms of time required for data analysis. For this comparison, only the dataset with the fewest scans was manually analyzed, with the time for other datasets estimated based on the assumption that each scan takes about one minute to analyze. Under this assumption, the analysis of the DRX dataset would take approximately 4.5 hours, not accounting for necessary breaks when dealing with large datasets. In contrast, the automated approach required only 5 minutes for each of the tested datasets, encompassing the time for generating training data, training the model, and analyzing the scans. Remarkably, these times were achieved using consumer-grade hardware[1]. Consequently, the automated approach not only demonstrates a significant reduction in analysis time compared to manual methods but also maintains a high level of accuracy, approximately 98%, highlighting its effectiveness.

Furthermore, the efficiency of the automated analysis framework can be enhanced by training the model concurrently with conducting the experiments. Since the synthesis and subsequent analysis of samples span several hours, a neural network

[1] CPU: AMD Ryzen 5 3600, RAM: 32 GB DDR4-3200, GPU: NVIDIA GTX1070

model can easily be trained in this time frame. Once the scans are obtained, they can be analyzed in seconds, thus, substantially reducing the time reported for employing the framework. This efficiency opens the possibility of integrating this novel framework into existing material discovery systems. These systems, typically only equipped with a CPU, could effectively run the model training while the samples are being synthesized. This is only possible thanks to the determination of an optimized model architecture, which minimizes computational efforts for training, as detailed in Chapter 4. Consequently, the proposed concept can be seamlessly integrated into existing systems without necessitating additional powerful hardware, thus enhancing these systems for fully automated material exploration experiments.

Additionally, the estimated times for manual evaluation of the datasets assume constant availability of personnel for analysis. In contrast, integrating the automated analysis approach into a system enables the material discovery platforms to conduct experiments based on the automated analysis of samples from earlier experiments without human intervention. For instance, in *in situ* instruments, where experiments are monitored in real-time, the system could strategically manipulate the temperature or conclude data acquisition based on the immediate feedback from the automated analysis. Similarly, in studies exploring the synthesis of a target structure, this feedback loop could guide the selection of precursor combinations and experimental conditions. Consequently, such systems are capable of running experiments fully autonomously every day during the week, greatly enhancing efficiency by removing the bottleneck of manual pattern analysis. This approach not only streamlines the experimental process but also maximizes the use of experimental time, leading to potentially faster and more efficient material discovery.

The application of the framework across different datasets has highlighted that novel materials are often challenging to synthesize. The training data generation approach presented in this thesis addresses this challenge by integrating a simulation tool for accurate computation of the diffraction patterns that represent the target structure. Consequently, in the context of material discovery experiments, the use of simulated signals for model training is without alternative, as measured

signals that contain the patterns of phase-pure samples are typically not available. This approach, however, is currently limited mainly to the generation of artificial XRD patterns due to the lack of simulation tools for Raman spectra. Nevertheless, the framework is designed to be adaptable; as soon as tools for rapid and accurate simulation of Raman spectra become available, they can easily be incorporated into the training data generation pipeline. This expansion would enable the application of this novel data analysis approach to Raman spectral analysis. Thus, this chapter showcases the high level of performance of the presented concept, emphasizing not only its accuracy and time efficiency but also its versatility in analyzing patterns from various domains, including different material systems, radiation wavelengths, instruments, and characterization techniques. This adaptability underscores the potential for the framework to be a powerful tool in diverse areas of material analysis.

# 7 Conclusion

Modern high-throughput material discovery platforms play a crucial role in the rapid discovery of novel materials, aiming to enhance the properties of existing substances in various applications. These platforms facilitate the automatic mixing of various precursors and enable the exploration of different material structures under a range of experimental conditions. The primary objective in these experiments is usually the successful synthesis of a specific target structure. This target structure is frequently predicted using advanced computational tools or identified within existing phase diagrams that describe the material system under investigation. However, it is important to note that not all combinations of precursors and experimental conditions successfully lead to the formation of the desired structures. Thus, while high-throughput platforms accelerate the production of samples, it is still necessary to investigate the resulting materials.

Sample characterization techniques like X-ray diffraction (XRD) and Raman spectroscopy are crucial for identifying crystalline phases formed during experiments, enabling the recognition of samples that match the target material based on their unique "fingerprint". However, the analysis of XRD patterns and Raman spectra is time-intensive and demands considerable expertise to determine the materials present in the produced samples accurately. Consequently, this manual identification process creates a substantial bottleneck in the workflow of existing high-throughput platforms, impeding the rapid discovery of novel materials that these platforms aim to achieve.

To address this challenge, a comprehensive framework was developed for the automated identification of novel materials in the experimental data. This framework automatically identifies samples whose characterization patterns align with

the characteristic fingerprint of the target structure. The core of this concept involves the generation of an extensive synthetic dataset that accurately depicts the diverse range of patterns found in material discovery datasets, either from XRD or Raman spectroscopy analysis. These patterns reflect a range of experiment outcomes, from successful to failed attempts, and are rapidly generated based on a detailed description of the target material.

Additionally, the framework features a neural network architecture that has been specifically tailored for the precise analysis of XRD patterns and Raman spectra. This architecture achieves an optimal balance between predictive accuracy and computational efficiency, ensuring that the model can be trained quickly and efficiently, even on hardware with limited computing power. Moreover, the framework includes methods that facilitate the training and application of these neural network models without necessitating specialized expertise in deep learning. This aspect significantly enhances the accessibility and usability of the framework, allowing users from various backgrounds to effectively employ this advanced analytical tool in their material discovery endeavors.

Crucially, the effectiveness of this framework is demonstrated on various experimental datasets, encompassing different material compositions and characterization techniques. When compared to manual data analysis, the neural network model exhibits almost identical accuracy but substantially reduces the time necessary for analyzing large datasets. Consequently, this framework offers a versatile solution for the automated analysis of diverse material discovery datasets. Its integration into existing high-throughput platforms can significantly expedite the data analysis process, thereby resolving the bottleneck of manual analysis and unlocking the potential for more efficient material discovery experiments.

Thus, the major contributions in this thesis can be summarized as follows:

1. The development of a novel concept integrating a neural network model for the automated analysis of data from material discovery experiments. Utilizing the proposed framework, neural networks can be trained and employed to identify targeted material structures in both XRD patterns and Raman spectra. The integration of this concept holds the potential to accelerate material discovery processes substantially, enhancing the efficiency of high-throughput pipelines (see Chapter 2).

2. The design of a robust data simulation framework that facilitates the training of neural networks. This method represents a substantial improvement over existing data simulation approaches, as it eliminates the need for manual identification of all possible materials and substances that can occur in the investigated material system. Accordingly, the required expertise to integrate such automated analysis approaches is reduced considerably. Since the training data generation is built around established simulation tools, it is currently only available for XRD pattern generation (see Chapter 3).

3. The introduction of an optimized neural network structure that exhibits high accuracy in analyzing XRD patterns and spectra with default properties. This innovative model architecture adeptly balances accuracy with model complexity, enabling fast training and application even on systems without high-performance computing capabilities. Despite this, it maintains an exceptional level of predictive quality, and consistently identifies the characteristic fingerprint of the target materials in the signals (see Chapter 4).

4. The implementation of the presented methods in the established programming language Python. This enabled leveraging its extensive library ecosystem, which includes several indispensable libraries for efficient training data generation and neural network model implementation and training. The code repositories have been made publicly available, enhancing the accessibility of the proposed data analysis concept. Its accessibility allows users to apply the framework for experimental data analysis without needing specialized knowledge in neural network design or training. Accordingly,

this implementation streamlines the process of integrating these tools into existing material discovery workflows (see Chapter 5).

5. The demonstration of successful analysis of datasets containing XRD patterns from material discovery experiments using the presented framework. The approach's versatility is evident, as it adeptly handles datasets with diverse properties and unique materials without necessitating any modifications to the framework. This high level of adaptability is further underscored by a comparative analysis against the manual analysis of the measured data, showing that the automated system achieves a prediction accuracy of approximately 98%, while significantly accelerating the analysis process (see Chapter 6).

6. The development of a general concept for evaluating distinct neural network models in the analysis of spectra with diverse properties. This advancement facilitates the comparison of various network architectures and can also be extended as more networks are developed. Therefore, it can be assured that alternative network architectures are indeed better than those presented in previous studies (see Section 4.3.2 and Section 5.3).

7. The design of a training data generation pipeline is notably adaptable, allowing for the future inclusion of data simulation tools for other characterization techniques as they become available. At present, the framework primarily utilizes simulation tools for computing XRD patterns. However, its structure is specifically crafted for adaptability, ensuring that new simulation tools can be seamlessly integrated as they are developed. This foresight in design means that the scope of the novel framework's application can be expanded in the future, accommodating a broader range of characterization techniques and thereby enhancing its utility in the field of material science. (see Section 3.5 and Section 5.4).

By employing the presented framework, it becomes feasible to design a high-throughput platform that utilizes the results of automated data analysis to inform and plan subsequent experiments in a feedback loop directly. Such platforms are theoretically capable of exploring the synthesis of novel materials by automatically adjusting precursors and experimental parameters in response to previous outcomes. Moreover, for platforms that monitor the formation of different structures in real-time, there is the potential to strategically manipulate variables such as the temperature to investigate the development of the target material further. Accordingly, the integration of this innovative framework is a crucial step towards creating more advanced material discovery platforms. These enhanced platforms could operate continuously, eliminating the downtime typically required for expert analysis of measured data between experiments. This continuous operation not only accelerates the material discovery process but also ensures a more efficient and seamless workflow, paving the way for rapid advancements in material science research.

In addition to XRD and Raman spectroscopy, material discovery platforms are typically equipped with a broad range of instruments to analyze the produced samples. For example, scanning electron microscopy (SEM) is used for the visual analysis of particles in the samples, providing insights into the size and shape of the particles, which are crucial factors influencing the material's properties. Therefore, studies are increasingly focusing on automating the analysis of such characterization methods [102]. To realize fully-automated workflows, automating the data analysis process for all integrated sample characterization techniques is crucial. Currently, limitations in analyzing various types of material discovery data hinder the full automation of some high-throughput platforms. Therefore, there is a significant need for advancements in automating data analysis across diverse domains to facilitate the complete automation of these platforms

# A Appendix

## A.1 Phase Identification using QualX

Several programs are available, both commercially and under academic licenses, for phase identification in XRD scans, utilizing the Figure of Merit (FoM) metric. Here, the tool QualX is employed as an illustrative example [39]. In addition, other programs like HighScore [40] and DIFFRAC.EVA [41] offer similar capabilities, each distinguished by its unique features and implementation of the FoM metric. These tools identify phases by comparing the measured diffraction patterns against reference patterns from various databases. Similar to the data analysis tools, a range of databases is available, such as the COD [35] and the ICSD [36], that provide this essential reference information. These databases are accessible under different models, ranging from free-to-use to those requiring commercial licenses.

The QualX program integrates the free POW_COD database for phase identification tasks [39]. Nonetheless, it is essential to highlight that the data analysis example presented here is based on the use of these two tools, and employing different programs or databases might yield varied results. In this instance, an XRD scan of Halite from the RRUFF database [28] is utilized to demonstrate the phase identification procedure using the FoM metric. As mentioned in Section 1.3.2, the scan predominantly shows peaks corresponding to the Halite structure's plane distances $d_{hkl}$. However, it also contains two additional peaks that are not attributable to Halite. Consequently, the phase identification process aims not only to identify the primary phase, Halite, but also to detect and provide insights into the impurity present in the sample.

Figure A.1 displays the graphical user interface (GUI) of the QualX application, showcasing the analysis of the XRD scan for the Halite sample. The GUI is divided into two primary sections: the upper part presents the visualization of the XRD scan under analysis, while the lower section displays the corresponding entries from the database. As outlined in Section 1.4.2, these data analysis tools calculate the theoretical positions and intensities of reflections based on the reference data for phase matching. Subsequently, the theoretical diffraction peak information is aligned with the measured data, and the reference phases are ranked according to the quality of their match using the FoM metric.

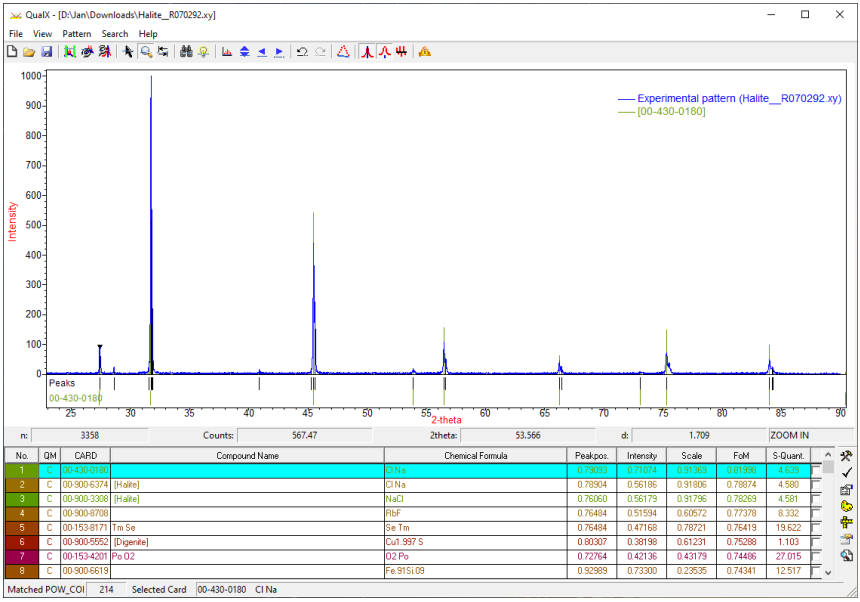



**Figure A.1:** Screenshot of the program QualX 2.0, which is commonly applied for analysis of XRD data. In the upper part of the layout, the measured intensity values are displayed, and the detected peaks are illustrated as black bars below the pattern. Below the illustration of the measured pattern is a list of phases from a connected database that have been matched to the measured values. The program orders the database entries by their quality of match, as quantified by the FoM metric. For visual assessment of the match, the calculated positions of the candidates are additionally visualized as green bars in the upper part, underneath the detected peak positions (black).

QualX incorporates the preprocessing routines described in Section 1.4.1 to extract crucial information on the position and intensities of peaks from the measured scan for pattern matching. Hence, to generate the filtered scan shown in Figure A.1, parameters related to smoothing, baseline removal, and peak search methods required fine-tuning. The positions of detected peaks are indicated by black bars beneath the blue signal. For this particular scan, QualX identified several candidate phases (No. 1-8) with high FoM values ranging from 0.82 to 0.74. The top-ranked entry, with the highest quality of fit, is listed under database card number 00-430-0180. Although it does not include the compound name, it is identified by the chemical formula Cl Na, corresponding to the evaluated sodium chloride sample. The second and third entries also display the same formula and also provide the name for the mineral: Halite. Consequently, the phase identification process strongly suggests that the Halite phase is present in the scanned sample.

The program additionally displays theoretical peak positions for visual confirmation of the match, shown as green bars below the extracted peak information, as demonstrated for entry No. 1. However, candidates No. 4-8 also achieved a FoM marginally lower than the correct Halite entries. These entries are characterized by a crystal structure with lower symmetry, which should result in a more complex diffraction pattern, exhibiting additional peaks in the range from 20° and 90°$2\theta$. In practice, an expert can manually discard these candidates by comparing the theoretical and measured peak positions. To exclude these samples based solely on FoM values, the metric's calculation formula needs adjustment. For instance, QualX's FoM implementation includes a penalty term for missing peaks in the measured pattern, and increasing this term can effectively differentiate entries No. 4-8 from the more accurate sodium chloride entries.

The diffraction pattern, in addition to the major peaks, reveals two additional peaks at approximately $29^\circ$ and $41^\circ 2\theta$ that are not attributable to the Halite structure. However, identifying the impurity phase in this sample solely based on this information is challenging, as no structure exclusively aligns with these two peak positions. Nonetheless, the corresponding impurity phase may produce other reflections in the pattern that either overlap with Halite peaks or are obscured

by noise in the data, thus rendering them undetectable. In practice, accurately determining the impurity phase is feasible only if the range of possible candidates is narrowed down using supplementary information. For instance, employing an alternative characterization technique to determine the predominant elements in the compound could limit the database search to phases containing only those elements, thereby facilitating more precise identification.

In addition to the major peaks, the diffraction pattern shows two additional peaks at about $29°$ and $41°2\theta$, which cannot be attributed to the Halite structure. However, determining the impurity phase in this sample is not possible based on the prevalent information alone, as there is no structure that only has peaks at these two positions. Alternatively, the additional peaks could overlap with the Halite peaks or are obstructed by the noise. In practice, it is only possible to determine the impurity phase if the list of candidates can be constrained using additional information. For example, an alternative characterization technique could be used to determine the prevalent elements in the compound so that the database is restricted to phases that only contain these elements. Consequently, while phase identification using the FoM metric is a valuable tool for discerning present phases, its accuracy often depends on the expertise of the user in fine-tuning preprocessing parameters, modifying the FoM calculation equation, or incorporating supplemental information to narrow down the list of potential candidates.

## A.2 Full Profile Analysis using GSAS-II

In full profile analysis, the Rietveld model is employed to approximate a measured signal based on a given crystal structure. To compare this approach with the FoM method, the identical XRD pattern of the Halite mineral from the RRUFF database [28] is analyzed. Figure A.2 displays a screenshot of the GSAS-II software [103], which fits a crystal structure of the corresponding phase, obtained from the ICSD (entry 29929), to the measured intensities. Using the Rietveld refinement approach, a full diffraction pattern is simulated, capturing peak positions and intensities, as well as the experimental artifacts of the measured scan.

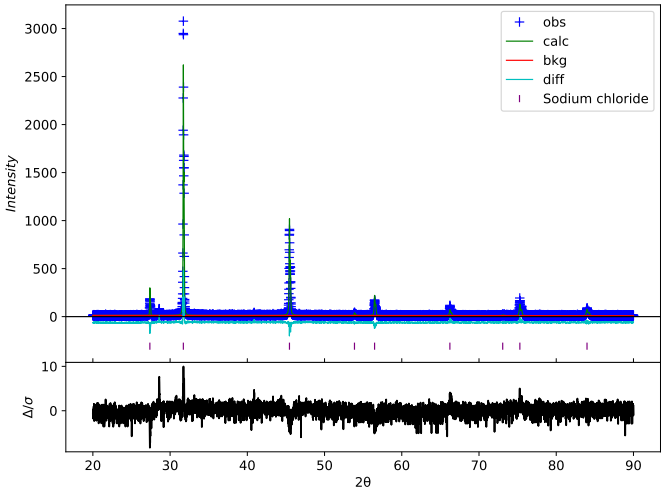



**Figure A.2:** A measurement of the mineral Halite (sodium chloride) is analyzed (from the RRUFF database [28]) and compared to a simulated pattern generated for the corresponding crystal structure (ICSD [36] entry 29929) using the GSAS-II software [103]. The software utilizes the Rietveld refinement method [104] to match the measured intensities with a model of the underlying crystal structure. The blue crosses present the observed intensity values, and the green plot illustrates the fitted pattern. Below, the purple bars demonstrate the computed diffraction positions for the refined structure, and the cyan line depicts the residual between simulated and measured intensity values. Underneath, the remaining differences ($\Delta$) are displayed and scaled according to the standard deviation ($\sigma$) of the measurement.

Consequently, both the cyan and black lines depict the differences between measured and simulated intensity values. While the cyan line shows the absolute differences, the black line scales the residual according to the standard deviation of the measured pattern ($\Delta/\sigma$), enhancing the interpretability of the fitted model.

Here, the remaining differences appear to constitute mostly the noise present in the scan, but for some angles, the residual stands out from the noise, e.g., for 29 and 32 deg. As indicated in Figure 1.8, the 29° peak cannot be attributed to the Halite structure, so it can be assumed that the sample contains impurity phases.

Here, a GoF metric of 1.41 was achieved for the measured XRD scan and the corresponding crystal structure, which indicates a good fit (a value of 0 means perfect fit). However, several steps were necessary to adjust the parameters of the Rietveld model and achieve a high-quality fit. For once, the lattice parameters of the evaluated mineral differ slightly from the database reference (5.6409 nm instead of 5.6338 nm), so a refinement of the unit cell was required to match the diffraction peak positions in the measured signal. Furthermore, an appropriate model to approximate the baseline intensities had to be selected, and a refinement of the parameters that approximate the peak profiles was necessary. Nonetheless, the selection of the appropriate parameters to include during the refinement varies between signals, so manual intervention is typically required to apply the Rietveld refinement.

## A.3 Neural Networks

Neural networks are mathematical models that form an integral part of the machine learning domain, specifically within the realm of deep learning. Machine learning is a subset of artificial intelligence that focuses on developing algorithms and models to solve specific tasks involving the analysis of data. In particular, the mathematical models are not explicitly programmed but rather learn to solve tasks over time as they are exposed to samples of the data. Deep learning, on the other hand, is a subfield of machine learning that focuses on the development and training of neural networks. These networks are modeled according to the structural and functional attributes observed in the neural architecture of the human brain (or other animals) responsible for processing sensory input [66, 105].

In practice, these models can automatically learn to identify complex features and patterns in the data, making them particularly effective for tasks such as image recognition and natural language processing. While alternative machine learning models such as Support-Vector Machines [106] or Random Forests [107] can also be effective for certain tasks, they typically require a manual feature extraction or transformation step, which can be time-consuming and require human expertise.


**(a)** Artificial neuron with inputs $x_n$, weights $w_n$, the bias $b$ and the activation function $\sigma$.

**(b)** Subsequent stacks of parallel neurons (layers) form a multi-layer perceptron.

**Figure A.3:** Schematic arrangement of neural networks and the basic neuron unit.

In contrast, deep learning algorithms are designed to automatically learn complex features and patterns in the raw data, eliminating the need for manual feature engineering [66]. Hence, neural networks typically excel when trained on massive datasets, allowing them to discover intricate patterns and relationships, while traditional machine learning algorithms tend to outperform neural networks when dealing with smaller datasets [108].

At a basic level, a neural network processes a given input $\boldsymbol{x}$ to produce a predicted output $\hat{y}$ through a series of mathematical operations, as summarized by function $f^*$ [66]. The basic unit of a neural network is a neuron, which takes the input $\boldsymbol{x}$, multiplies it by a set of weights $\boldsymbol{w}$, adds a bias $b$, and then applies an activation function $\sigma$ to produce an output, as depicted in Figure A.3a. Mathematically, this can be expressed as

$$z = \boldsymbol{x} \cdot \boldsymbol{w} + b, \tag{A.1}$$

where $z$ is the weighted sum of the input. The output of the neuron, here denoted as $\hat{y}$, is obtained by applying the activation function [66]:

$$\hat{y} = \sigma(z). \tag{A.2}$$

In a single layer, multiple neurons operate in parallel, each having its own set of weights and biases. The outputs of these neurons form the layer's output vector $\boldsymbol{o}$. This layer-wise computation is a fundamental building block, and neural networks extend this concept by stacking multiple layers on top of each other. The output of one layer becomes the input of the next, creating a hierarchical structure, as

conceptualized in Figure A.3b. Mathematically, for a given layer $l$, the output $\boldsymbol{o}^{(l)}$ can be expressed as:

$$\boldsymbol{o}^{(l)} = \sigma(\boldsymbol{W}^{(l)} \cdot \boldsymbol{o}^{(l-1)} + \boldsymbol{b}^{(l)}), \tag{A.3}$$

where $\boldsymbol{W}^{(l)}$ represents the matrix of weights, $\boldsymbol{o}^{(l-1)}$ is the output from the previous layer, $\boldsymbol{b}^{(l)}$ is the vector of biases for the current layer, and $\sigma$ is the activation function[66]. This layer-wise connectivity allows neural networks to learn intricate representations from the input data, enabling them to capture complex patterns and relationships. The final prediction $\hat{y}$ is obtained from the output of the last layer. As a result, $f^*$ encapsulates the entire process of transforming the input $\boldsymbol{x}$ through the layers of the neural network to produce the final prediction $\hat{y}$.

To enable the network to learn an represent complex patterns in the data, activation functions are used within the layers of a neural network to introduce non-linearities. The introduction of non-linear activation functions is crucial because a neural network composed solely of linear operations, such as simple matrix multiplications and additions, would effectively behave like a single-layer perceptron [66]. In such cases, regardless of the network's depth, the overall transformation would remain linear. Commonly used activation functions in neural networks are, for example, the sigmoid activation function $\sigma$

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{A.4}$$

and the Rectified Linear Unit (ReLU)

$$\text{ReLU}(z) = \max(0, z). \tag{A.5}$$

Neural networks offer a versatile framework capable of tackling a multitude of tasks, with two of the most fundamental being regression and classification. In regression tasks, the goal is to predict a continuous output, such as predicting the price of an object based on historic sales data. In this scenario, the network's output layer typically consists of a single neuron, and the predicted value $\hat{y}$ can take any real number. To model the output to fit the desired range of values, activation functions like linear ($\sigma(z) = z$) or ReLU can be considered. On the other hand, in classification tasks, the objective is to categorize input data $\boldsymbol{x}$ into distinct classes $c_n$. Here, the selection of an appropriate activation function depends on the nature of the task. For tasks where input samples exclusively belong to one class (multi-class), the softmax activation function is commonly employed

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}. \tag{A.6}$$

Softmax scales the values of the neurons in the network's final layer $z_1, \ldots, z_n$ into a probability distribution, ensuring that the sum of probabilities across all classes equals one. In this case, the corresponding class for a given input is determined by identifying the class with the highest probability. Conversely, in multi-label tasks where an input can belong to multiple classes simultaneously, the sigmoid activation function is preferred. Thus, the sigmoid activation independently scales the output for each class between 0 and 1, providing a probability estimate for the presence of each class. In the case of sigmoid activation, a predicted probability $\hat{y}_i > 0.5$ is indicative of the presence of the corresponding class $c_i$.

For a given input $\boldsymbol{x}$, a neural network with underlying parameters $\Phi$ (including $\boldsymbol{W}, \boldsymbol{b}$) provides a prediction $\hat{\boldsymbol{y}}$. However, to ensure that the predicted output $\hat{\boldsymbol{y}}$ aligns with the actual output $\boldsymbol{y}$, the model's parameters first have to be adjusted in a training process. In the realm of supervised learning, the essence lies in the availability of input-output pairs that represent the relationship between the input data $\boldsymbol{x}$ and its corresponding label or output $\boldsymbol{y}$. Provided the desired output for a given input is known, one can compare the prediction of the network $\hat{\boldsymbol{y}}$ with

the label to approximate the modifications required to the model's parameters to align the prediction with the label. This comparison is facilitated by the use of a designated metric known as a loss function. The loss function quantifies the dissimilarity between the predicted output and the true label, providing a numerical measure of the model's performance. The objective during training is to minimize this loss, prompting adjustments to the model's parameters $\Phi$ through optimization techniques [66]. For a training set containing $n$ samples, the loss $L$ can be described as

$$L = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{\boldsymbol{y}}_i, \boldsymbol{y}_i), \tag{A.7}$$

where $\ell$ describes the loss function. Depending on the task and activation function in the output layer, different loss functions can be used. For example, the mean squared error (MSE), which is commonly used for regression tasks and linearly scaled outputs, is calculated as follows:

$$\ell_{\text{MSE}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = (\hat{\boldsymbol{y}} - \boldsymbol{y})^2. \tag{A.8}$$

For classification tasks, on the other hand, the network typically predicts probability estimates for $M$ outputs, so the cross-entropy loss (CE) is used [66]

$$\ell_{\text{CE}}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{j=1}^{M} y_j \log(\hat{y}_j). \tag{A.9}$$

To minimize the loss during the training process, the model's parameters $\Phi$ are adjusted to align prediction and label. This algorithm calculates the gradient of the loss function with respect to each weight and bias in the network. The gradient indicates the direction of the steepest ascent, so adjusting the weights and biases in the direction of the negative gradient reduces the error (loss function) of the model [66].

Because of limited available memory during training, neural networks are trained using batches of data instead of the entire dataset at once, enhancing computational efficiency. The pace of the iterative weight updates is governed by a predetermined learning rate, allowing the model to converge gradually. By using the learning rate, the model can avoid overshooting the optimal solution, allowing for a controlled and gradual convergence towards the minimum of the loss function. Although training with batches may result in noisier weight updates, potentially helping the model escape local minima, it can also make the convergence process less smooth, which results in extended training times or a model that fails to fit properly. Thus, optimizers, such as the Stochastic Gradient Descent (SGD), Adam (derived from adaptive moment estimation), or Root Mean Square Propagation (RMSprop), are used to enhance the gradient descent approach. Optimizers introduce advanced concepts, including momentum or adaptive learning rates, which facilitate a balance between steady convergence and the ability to escape local minima, leading to better generalization in the model [66].

The universal approximator property of neural networks means they can theoretically model any continuous function, given enough hidden units. This flexibility, while powerful, can lead to overfitting when the network is too complex for the data. Overfitting occurs when a network starts to memorize the noise present in the training data instead of learning the actual underlying patterns, which diminishes its ability to generalize effectively to new, unseen data. Alternatively, specialized layers can be incorporated into the neural networks that restrict the parameters in the network and therefore help to prevent overfitting. For example, convolutional layers utilize a kernel with shared weights that is shifted across the input, enhancing the network's capability to identify local features that are invariant to position. Furthermore, by hierarchically combining these local features through multiple convolutional layers, the network can learn to recognize increasingly complex patterns within the inputs [66]. This approach has led to the widespread adoption of convolutional layers in various fields that utilize neural networks for automated data analysis, including image recognition, where they have proven to be particularly effective [84, 85].

In addition to reducing the number of weights through the use of convolutional kernels that slide over the input, regularization methods are used in neural networks to prevent overfitting. By adding constraints to the network's complexity, regularization techniques ensure that the model generalizes well from the training data to new, unseen data. Common methods include L1 and L2 regularization, which add penalties to the loss function based on the weights' magnitudes, and dropout, where random units in a layer are "dropped out" or set to zero during training, forcing the network to spread out its learning across the weights. These methods help in creating a model that is complex enough to fit the data well but not so complex that it fits the noise in the training data, leading to better performance on unseen data [66].

Figure A.4 illustrates the functionality of a convolutional neural network (CNN) in the domain of image recognition. An image with two elephants serves as the input to a network that includes several convolutional layers, each containing multiple filters. In the network's early stages, the filters in these layers identify basic shapes including the elephants' outlines, the floor, or the sky, each represented on separate feature maps. As the data progresses through the network, the later convolutional layers detect more complex features and patterns, such as the elephants' trunks. Rather than increasing the size of the convolutional kernels to detect spatially extensive features, pooling layers are interspersed among the convolutional layers to reduce the dimensionality of the input, which simultaneously broadens the receptive field of the kernels that follow. Then, the fully-connected layers relate the feature maps with the classes and provide a numerical prediction for each class, which is transformed into probabilities through the softmax activation function [66].

State-of-the-art convolutional architectures often incorporate advanced concepts like batch normalization and complex stacks of the convolutional layers to enhance performance. Batch normalization [109] normalizes the activations within a layer, making the training process more stable and allowing for higher learning rates, thus speeding up convergence. By stacking more convolutional layers, as prominently introduced by the VGG architecture [85], the network is able to detect more complex patterns in the data, but deep networks suffer from

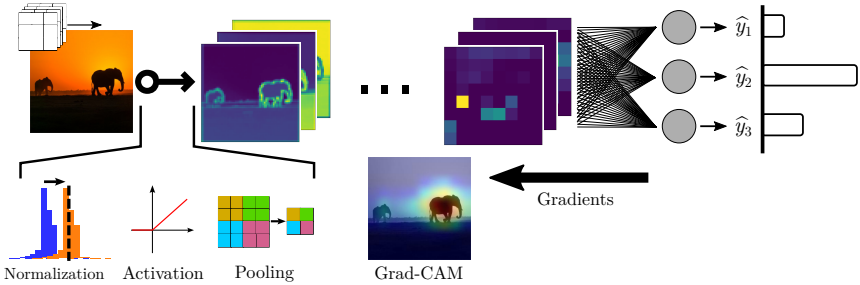



**Figure A.4:** Schematic functionality of a convolutional neural network (CNN), with multiple filters shifted across the input to identify feature maps. The typical structure includes batch normalization, activations to introduce non-linearities, and pooling layers. After the convolutional layers, fully-connected layers are applied to the complex feature maps and provide a classification output $\hat{y}_n$. Based on the gradients of the network, the most relevant regions of the input are calculated and shown as a heatmap for the exemplary input (Grad-CAM).

vanishing gradients, so the weights cannot be adjusted if the network contains too many layers. The VGG architecture, in particular, stacks multiple convolutional layers prior to the pooling operations. Alternatively, residual blocks, introduced in architectures like ResNet [84], include skip connections that bypass one or more layers, making it easier to train very deep networks by mitigating the vanishing gradient problem. Additionally, inception blocks [86] use kernels of different dimensions to detect features of various sizes, leading to networks with a broader range of capabilities [110].

In Figure A.4, an Xception network [111], which is an advancement of the inception architecture that includes batch normalization and skip connections, was used to analyze the image with two elephants. The early convolutional layers in CNNs, including advanced architectures like the Xception model, typically focus on detecting simple features such as outlines and basic shapes. However, the interpretation of detected features becomes more complex in the later stages of the deep neural networks, as the layers encode higher-level and more abstract features.

To gain insights into what these later layers are focusing on, visualization techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) can be

used, which highlight the important regions in the image for making a particular classification decision [112]. Here, the Grad-CAM in Figure A.4 reveals that features of the larger elephant are more influential for the classification, due to the greater gradients to this region of the input. Nonetheless, the gradients related to the region of the smaller elephant are greater than those of the background, so even this region contributes to the final prediction of the network. This highlights the functionality of the convolutional layers, in which filters are shifted across the input, so matching features of both elephants are identified concurrently.

## A.4 Python Package Usage

### A.4.1 python-powder-diffraction

**Listing A.1:** Minimal working example to utilize the Powder class.

```
from powdiffrac import Powder
powder = Powder.from_cif(
  "structure.cif", # full path to cif file
  #arguments concerning the scans
  two_theta: tuple = (10,80), #2t range (Min, Max)
  step_size: float = 0.01, #step size of scans
  #arguments controlling the variations
  max_strain: float = 0.04, #strain on the lattice
  max_texture: float = 0.6, #texture limit
  min_domain_size: float = 10, #grain size limit
  max_domain_size: float = 100, #grain size limit
  #arguments to switch on/off the variations
  vary_strain: bool = False,
  vary_texture: bool = False,
  vary_domain: bool = False,
)
signal = powder.get_signal()
```

**Listing A.2:** Minimal working example to utilize the noise generation function

```
from powdiffrac.simulation import generate_noise
#assume x is numpy array containing the signals
#in form [num, length]
#add noise to x and save as x_noise
x_noise = generate_noise(
  x, # input array
  #random seed
  seed=None,
  #chebyshev polynomials
  cheb=None, #None -> random polynomials
  #level of noise in scans
  noise_lvl=None, #None -> draw from interval
  noise_min=0.01,
  noise_max=0.03,
)
```

**Listing A.3:** Example for running training data generation from command-line interface. For each cif in the folder "all_cifs", 120 XRD patterns are generated. The patterns are simulated in the range from 10 to 70 degrees $2\theta$ with a step size of 0.02° $\Delta 2\theta$. The grain size is provided explicitly in the interval [20,50). There is no preferred orientation simulation included in the scans (texture), but strain and domain size variation are "on" by default.

```
generate_training_data "./all_cifs/" -theta_range "(10,70)"
-step_scan 0.02 -domain_sizes "(20,50)" -no_var_texture
-n_train 100 -n_val 20
```

### A.4.2 spectra-network-benchmark

**Listing A.4:** Adaptable parameters in the script to generate synthetic signals.

```
# snippet from generate_dataset file

# GLOBAL PARAMETERS
signal_length =  5000
shift_range = 100 # all peaks shifted independently
variation_range = 0.1 # +/- absolute height for each peak
kernel_range = (2, 30)

# Target Fingerprint
positions_phase = [4290, 4700]
heights_phase = [0.744, 1.0]

# Alternative Fingerprints
min_peaks = 2
max_peaks = 10

n_train = 5000
```

### A.4.3 crystal-id

**Listing A.5:** Exemplary config.yml file to simulate training data. The file specifies that diffraction patterns are simulated with a copper K$\alpha$ wavelength. The scans are simulated between 10 and 100 $2\theta$ with a step size of 0.01° $\Delta 2\theta$. In total, 1000 patterns are simulated.

```
domain:
  method: XRD
  radiation: CuKa # alternative wavelength as float

measurement:
  range: [10.0, 100.0]
  step: 0.01
  holder_position: null
  impurity_cutoff: 0.0

simulation:
  n_patterns: 1000
  lattice_variation: 0.02
  intensity_variation: 0.05
  fwhm: 30
  background_ratio: 0.1
  noise_ratio: 0.03
```

**Listing A.6:** Example for running training data generation and subsequent model training from command-line interface. For a provided material structure in the form of a cif, varied signals are generated. The patterns are simulated according to the values in the corresponding config.yml file. Here, exemplary signals for the material system "halite" are generated. The input size of the model is automatically adapted to the length of the simulated scans.

```
generate_training_data "halite"
-structure_path "./halite.cif"

train "halite"
```

# List of Figures

# List of Tables

# List of Publications

## Journal articles

[68] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, "Enhancing deep-learning training for phase identification in powder x-ray diffractograms," in *IUCrJ*, vol. 8, pp. 408 – 420, 2021.

[83] J. Schuetzke, N. J. Szymanski, and M. Reischl, "Validating neural networks for spectroscopic classification on a universal synthetic dataset," in *npj Computational Materials*, vol. 9, no. 1, p. 100, 2023.

[67] J. Schuetzke, S. Schweidler, F. R. Münke, A. Orth, A. D. Khandelwal, B. Breitung, J. Aghassi-Hagmann, and M. Reischl, "Accelerating materials discovery: Automated identification of prospects from XRD data in fast screening experiments," in *Advanced Intelligent Systems*, vol. 6, no. 3, p. 2300501, 2024.

[113] S. Sheshachala, B. Huber, J. Schuetzke, R. Mikut, T. Scharnweber, C. M. Domínguez, H. Mutlu, and C. M. Niemeyer, "Charge controlled interactions between dna-modified silica nanoparticles and fluorosurfactants in microfluidic water-in-oil droplets," in *Nanoscale Advances*, vol. 5, no. 15, pp. 3914–3923, 2023.

[108] F. R. Münke, J. Schuetzke, F. Berens, and M. Reischl, "A Review of Adaptable Conventional Image Processing Pipelines and Deep Learning on limited Datasets," in *Machine Vision and Applications*, vol. 35, no. 2, p. 25, 2024.

[102] L. Rettenberger, N. J. Szymanski, Y. Zeng, J. Schuetzke, S. Wang, G. Ceder, and M. Reischl, "Uncertainty-aware particle segmentation for electron microscopy at varied length scales," in *npj Computational Materials*, vol. 10, no. 1, p. 124, 2024.

[114] T. Phan-Xuan, S. Schweidler, S. Hirte, M. Schüller, L. Lin, A. D. Khandelwal, K. Wang, J. Schuetzke, M. Reischl, C. Kübele *et al.*, "Using the high-entropy approach to obtain multimetal oxide nanozymes: Library synthesis, in silico structure–activity, and immunoassay performance," in *ACS Nano*, vol. 18, no. 29, pp. 19 024–19 037, 2024.

## Conference contributions

[115] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, "Siamese Networks for 1D Signal Identification," in *Proceedings 30. Workshop Computational Intelligence, Berlin, Germany, November 26-27, 2020.* Karlsruhe, Germany: KIT Scientific Publishing, 2020, pp. 17–31.

[116] J. Schuetzke, B. Jones, N. Henderson, N. Rodesney, A. Benedix, K. Knorr, R. Mikut, and M. Reischl, "Application of Machine Learning to XRD Phase Identification," in *Denver X-Ray Conference*, Denver, CO, USA, 2020, DOI: https://dx.doi.org/10.5445/IR/1000127718.

[81] J. Schuetzke, N. J. Szymanski, and M. Reischl, "A Critical Review of Neural Networks for the Identification of Powder X-ray Diffraction Patterns," in *Advances in X-ray Analysis, Proceedings of the Denver X-ray Conference*, Rockville, MD, USA, 2022.

[82] J. Schuetzke, N. J. Szymanski, G. Ceder, and M. Reischl, "A Critical Review of Neural Networks for the Use with Spectroscopic Data," in *European Crystallographic Meeting (ECM)*, Versailles, France, 2022, DOI: https://dx.doi.org/10.5445/IR/1000151442.

# Bibliography


[1] T. Ahmad and D. Zhang, "A critical review of comparative global historical energy consumption and future demand: The story told so far," in *Energy Reports*, vol. 6, pp. 1973–1991, 2020.

[2] K.-S. Kim, S.-H. Lee, V. Q. Nguyen, Y. Yun, and S. Kwon, "Ablation characteristics of rocket nozzle using HfC-SiC refractory ceramic composite," in *Acta Astronautica*, vol. 173, pp. 31–44, 2020.

[3] A. I. Taub and A. A. Luo, "Advanced lightweight materials and manufacturing processes for automotive applications," in *MRS Bulletin*, vol. 40, no. 12, pp. 1045–1054, 2015.

[4] J. J. Hanak, "The "multiple-sample concept" in materials research: Synthesis, compositional analysis and testing of entire multicomponent systems," in *Journal of Materials Science*, vol. 5, pp. 964–971, 1970.

[5] P. Ghavami, *Mechanics of Materials: An Introduction to Engineering Technology*. Cham, Switzerland: Springer, 2015.

[6] E. W. McFarland and W. H. Weinberg, "Combinatorial approaches to materials discovery," in *Trends in Biotechnology*, vol. 17, no. 3, pp. 107–115, 1999.

[7] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: review of state of the art," in *ACS combinatorial science*, vol. 13, no. 6, pp. 579–633, 2011.

[8] R. Grainger and S. Whibley, "A perspective on the analytical challenges encountered in high-throughput experimentation," in *Organic Process Research & Development*, vol. 25, no. 3, pp. 354–364, 2021.

[9] N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim, and G. Ceder, "Toward autonomous design and synthesis of novel inorganic materials," in *Materials horizons*, vol. 8, no. 8, pp. 2169–2198, 2021.

[10] V. K. Pecharsky and P. Y. Zavalij, *Fundamentals of powder diffraction and structural characterization of materials*. New York, NY, USA: Springer, 2005.

[11] I. G. Clayson, D. Hewitt, M. Hutereau, T. Pope, and B. Slater, "High throughput methods in the synthesis, characterization, and optimization of porous materials," in *Advanced Materials*, vol. 32, no. 44, p. 2002780, 2020.

[12] C. R. Matos, M. J. Xavier, L. S. Barreto, N. B. Costa, and I. F. Gimenez, "Principal component analysis of x-ray diffraction patterns to yield morphological classification of brucite particles," in *Analytical chemistry*, vol. 79, no. 5, pp. 2091–2095, 2007.

[13] L. A. Baumes, M. Moliner, and A. Corma, "Design of a full-profile-matching solution for high-throughput analysis of multiphase samples through powder x-ray diffraction," in *Chemistry–A European Journal*, vol. 15, no. 17, pp. 4258–4269, 2009.

[14] C. J. Long, D. Bunker, X. Li, V. L. Karen, and I. Takeuchi, "Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization," in *Review of Scientific Instruments*, vol. 80, no. 10, p. 103902, 2009.

[15] H. Wang, Y. Xie, D. Li, H. Deng, Y. Zhao, M. Xin, and J. Lin, "Rapid identification of x-ray diffraction patterns based on very limited data by interpretable convolutional neural networks," in *Journal of Chemical Information and Modeling*, vol. 60, pp. 2004–2011, 2020.

[16] J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh, and K.-S. Sohn, "A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic xrd powder patterns," in *Nature Communications*, vol. 11, pp. 1–11, 2020.

[17] J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, "Deep convolutional neural networks for raman spectrum recognition: A unified solution," in *The Analyst*, vol. 142, no. 21, pp. 4067–4074, 2017.

[18] G. S. Rohrer, *Structure and bonding in crystalline materials*. Cambridge, UK: Cambridge University Press, 2001.

[19] T. Liavitskaya and S. Vyazovkin, "All You Need to Know about the Kinetics of Thermally Stimulated Reactions Occurring on Cooling," in *Molecules*, vol. 24, no. 10, p. 1918, 2019.

[20] C. Kittel, *Introduction to solid state physics*, 8th ed. Hoboken, NJ, USA: Wiley, 2005.

[21] C. C. Land, D. E. Peterson, and R. B. Roof, "Phase investigations of the Pu-Pt, Pu-Rh, and Pu-Pt-Rh systems," in *Journal of Nuclear Materials*, vol. 75, no. 2, pp. 262–273, 1978.

[22] W. Friedrich, P. Knipping, and M. Laue, "Interferenzerscheinungen bei Röntgenstrahlen," in *Annalen der Physik*, vol. 346, no. 10, pp. 971–988, 1913.

[23] W. C. Röntgen, "Ueber eine neue Art von Strahlen," in *Sitzungsberichte der Physik.-medic. Gesellschaft Wuerzburg*, Wuerzburg, Germany, 1895, pp. 137–147.

[24] W. L. Bragg and W. H. Bragg, "The reflection of x-rays by crystals," in *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 88, no. 605, pp. 428–438, 1913.

[25] A. Morawiec, *Indexing of Crystal Diffraction Patterns: From Crystallography Basics to Methods of Automatic Indexing*. Cham, Switzerland: Springer, 2022, vol. 326.

[26] K. D. M. Harris and P. A. Williams, "Powder Diffraction," in *Structure from Diffraction Methods*, D. W. B. Bruce, D. O'Hare, and R. I. Walton, Eds. Chichester, West Sussex, UK: John Wiley & Sons, Ltd, 2014, pp. 1–81.

[27] J. Brentano, "Focussing method of crystal powder analysis by X-rays," in *Proceedings of the Physical Society of London*, vol. 37, no. 1, pp. 184 – 193, 1924.

[28] B. Lafuente, R. T. Downs, H. Yang, and N. Stone, "The power of databases: The RRUFF project," in *Highlights in Mineral Crystallography*, T. Armbruster and R. M. Danisi, Eds. Berlin, Germany: De Gruyter, 2015, pp. 1–30.

[29] E. Smith, *Modern Raman spectroscopy: a practical approach*, 2nd ed., G. Dent, Ed. Hoboken, NJ, USA: Wiley, 2019.

[30] P. Colomban, "Analysis of strain and stress in ceramic, polymer and metal matrix composites by raman spectroscopy," in *Advanced engineering materials*, vol. 4, no. 8, pp. 535–542, 2002.

[31] C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. Saleh, S. Ermon, and J. Dionne, "Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning," in *Nature communications*, vol. 10, no. 1, p. 4927, 2019.

[32] T. W. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler, and J. Popp, "Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!?" in *Analytical Chemistry*, vol. 88, no. 1, pp. 133–151, 2016.

[33] J. De Gelder, K. De Gussem, P. Vandenabeele, and L. Moens, "Reference database of Raman spectra of biological molecules," in *Journal of Raman Spectroscopy*, vol. 38, no. 9, pp. 1133–1147, 2007.

[34] A. Wang, J. Han, L. Guo, J. Yu, and P. Zeng, "Database of standard Raman spectra of minerals and related inorganic crystals," in *Applied Spectroscopy*, vol. 48, no. 8, pp. 959–968, 1994.

[35] S. Gražulis, D. Chateigner, R. T. Downs, A. F. T. Yokochi, M. Quirós, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, and A. Le Bail, "Crystallography Open Database – an open-access collection of crystal structures," in *Journal of Applied Crystallography*, vol. 42, no. 4, pp. 726–729, 2009.

[36] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, "New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design," in *Acta Crystallographica Section B: Structural Science*, vol. 58, no. 3, pp. 364–369, 2002.

[37] P. C. A. M. Buijtels, H. F. M. Willemse-Erix, P. L. C. Petit, H. P. Endtz, G. J. Puppels, H. A. Verbrugh, A. van Belkum, D. van Soolingen, and K. Maquelin, "Rapid Identification of Mycobacteria by Raman Spectroscopy," in *Journal of Clinical Microbiology*, vol. 46, no. 3, pp. 961–965, 2008.

[38] Y. Iwasaki, A. G. Kusne, and I. Takeuchi, "Comparison of dissimilarity measures for cluster analysis of x-ray diffraction data from combinatorial libraries," in *npj Computational Materials*, vol. 3, no. 1, pp. 1–9, 2017.

[39] A. Altomare, N. Corriero, C. Cuocci, A. Falcicchio, A. Moliterni, and R. Rizzi, "*QUALX2.0* : a qualitative phase analysis software using the freely available database POW_COD," in *Journal of Applied Crystallography*, vol. 48, no. 2, pp. 598–603, 2015.

[40] T. Degen, M. Sadki, E. Bron, U. König, and G. Nénert, "The highscore suite," in *Powder diffraction*, vol. 29, pp. 13–18, 2014.

[41] Bruker AXS, "Diffrac.EVA: Software for the analysis of 1D and 2D X-ray datasets," 2023. URL: https://www.bruker.com/eva Last retrieved 2023-08-04.

[42] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, and K.-S. Sohn, "Classification of crystal structure using a convolutional neural network," in *IUCrJ*, vol. 4, pp. 486–494, 2017.

[43] N. J. Szymanski, C. J. Bartel, Y. Zeng, Q. Tu, and G. Ceder, "Probabilistic deep learning approach to automate the interpretation of multi-phase diffraction spectra," in *Chemistry of Materials*, vol. 33, no. 11, pp. 4204–4215, 2021.

[44] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. Tian, G. Romano *et al.*, "Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks," in *npj Computational Materials*, vol. 5, pp. 1–9, 2018.

[45] X. Fan, W. Ming, H. Zeng, Z. Zhang, and H. Lu, "Deep learning-based component identification for the raman spectra of mixtures," in *Analyst*, vol. 144, no. 5, pp. 1789–1798, 2019.

[46] P. M. Vecsei, K. Choo, J. Chang, and T. Neupert, "Neural network based classification of crystal symmetries from x-ray diffraction patterns," in *Physical Review B*, vol. 99, no. 24, p. 245120, 2019.

[47] P. C. Angelo and B. Ravisankar, *Introduction to steels: Processing, Properties, and Applications*. Boca Raton, FL, USA: CRC Press, 2019.

[48] X.-D. Xiang, X. Sun, G. Briceño, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, and P. G. Schultz, "A Combinatorial Approach to Materials Discovery," in *Science*, vol. 268, no. 5218, pp. 1738–1740, 1995.

[49] H. Koinuma and I. Takeuchi, "Combinatorial solid-state chemistry of inorganic materials," in *Nature materials*, vol. 3, no. 7, pp. 429–438, 2004.

[50] K. Kennedy, T. Stefansky, G. Davy, V. F. Zackay, and E. R. Parker, "Rapid method for determining ternary-alloy phase diagrams," in *Journal of Applied Physics*, vol. 36, no. 12, pp. 3808–3810, 1965.

[51] D. R. West and N. Saunders, *Ternary phase diagrams in materials science*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2017.

[52] G. Cacciamani, J. De Keyzer, R. Ferro, U. E. Klotz, J. Lacaze, and P. Wollants, "Critical evaluation of the Fe–Ni, Fe–Ti and Fe–Ni–Ti alloy systems," in *Intermetallics*, vol. 14, no. 10, pp. 1312–1325, 2006.

[53] L. Velasco, J. S. Castillo, M. V. Kante, J. J. Olaya, P. Friederich, and H. Hahn, "Phase–Property Diagrams for Multicomponent Oxide Systems toward Materials Libraries," in *Advanced Materials*, vol. 33, no. 43, p. 2102301, 2021.

[54] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," in *Physical Review B*, vol. 89, no. 9, p. 094104, 2014.

[55] K. Choudhary, I. Kalish, R. Beams, and F. Tavazza, "High-throughput identification and characterization of two-dimensional materials using density functional theory," in *Scientific reports*, vol. 7, no. 1, p. 5179, 2017.

[56] K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," in *npj Computational Materials*, vol. 7, no. 1, p. 185, 2021.

[57] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer *et al.*, "Graph neural networks for materials science and chemistry," in *Communications Materials*, vol. 3, no. 1, p. 93, 2022.

[58] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, "The open quantum materials database (oqmd): assessing the accuracy of dft formation energies," in *npj Computational Materials*, vol. 1, no. 1, pp. 1–15, 2015.

[59] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation," in *APL Materials*, vol. 1, no. 1, p. 011002, 2013.

[60] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, "Scaling deep learning for materials discovery," in *Nature*, vol. 624, pp. 80–85, 2023.

[61] N. J. Szymanski, Y. Zeng, T. Bennett, S. Patil, J. K. Keum, E. C. Self, J. Bai, Z. Cai, R. Giovine, B. Ouyang *et al.*, "Understanding the Fluorination of Disordered Rocksalt Cathodes through Rational Exploration of Synthesis Pathways," in *Chemistry of Materials*, vol. 34, no. 15, pp. 7015–7028, 2022.

[62] M. Bagheri and H.-P. Komsa, "High-throughput computation of raman spectra from first principles," in *Scientific Data*, vol. 10, no. 1, p. 80, 2023.

[63] Q. Liang, S. Dwaraknath, and K. A. Persson, "High-throughput computation and evaluation of raman spectra," in *Scientific data*, vol. 6, no. 1, p. 135, 2019.

[64] Z. Kou, A. Hashemi, M. J. Puska, A. V. Krasheninnikov, and H.-P. Komsa, "Simulating raman spectra by combining first-principles and empirical potential approaches with application to defective mos2," in *npj Computational Materials*, vol. 6, no. 1, p. 59, 2020.

[65] A. Kersch, R. Ganser, and M. Trien, "Simulation of xrd, raman and ir spectrum for phase identification in doped hfo2 and zro2," in *Frontiers in Nanotechnology*, vol. 4, p. 1026286, 2022.

[66] I. Goodfellow, *Deep learning*, ser. Adaptive computation and machine learning, Y. Bengio and A. Courville, Eds. Cambridge, Massachusetts: The MIT Press, 2016.

[67] J. Schuetzke, S. Schweidler, F. R. Münke, A. Orth, A. D. Khandelwal, B. Breitung, J. Aghassi-Hagmann, and M. Reischl, "Accelerating materials discovery: Automated identification of prospects from XRD data in fast screening experiments," in *Advanced Intelligent Systems*, vol. 6, no. 3, p. 2300501, 2024.

[68] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, "Enhancing deep-learning training for phase identification in powder x-ray diffractograms," in *IUCrJ*, vol. 8, pp. 408 – 420, 2021.

[69] P. Jahoda, I. Drozdovskiy, S. J. Payler, L. Turchi, L. Bessone, and F. Sauro, "Machine learning for recognizing minerals from multispectral data," in *Analyst*, vol. 146, no. 1, pp. 184–195, 2021.

[70] X. Sang, R. Zhou, Y. Li, and S. Xiong, "One-dimensional deep convolutional neural network for mineral classification from raman spectroscopy," in *Neural Processing Letters*, vol. 54, pp. 677–690, 2022.

[71] M. H. Mozaffari and L.-L. Tay, "Overfitting one-dimensional convolutional neural networks for raman spectra identification," in *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 272, p. 120961, 2022.

[72] A. Bhattacharya, J. A. Benavides, L. F. Gerlein, and S. G. Cloutier, "Deep-learning framework for fully-automated recognition of tio2 polymorphs based on raman spectroscopy," in *Scientific Reports*, vol. 12, no. 1, p. 21874, 2022.

[73] N. Ibtehaz, M. E. Chowdhury, A. Khandakar, S. Kiranyaz, M. S. Rahman, and S. M. Zughaier, "Ramannet: a generalized neural network architecture for raman spectrum analysis," in *Neural Computing and Applications*, vol. 35, pp. 18 719–18 735, 2023.

[74] A. K. Conlin, E. B. Martin, and A. J. Morris, "Data augmentation: an alternative approach to the analysis of spectroscopic data," in *Chemometrics and Intelligent Laboratory Systems*, vol. 44, no. 1, pp. 161–173, 1998.

[75] K. Georgouli, M. T. Osorio, J. Martinez Del Rincon, and A. Koidis, "Data augmentation in food science: Synthesising spectroscopic data of vegetable oils for performance enhancement," in *Journal of Chemometrics*, vol. 32, no. 6, p. e3004, 2018.

[76] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis," in *Computational Materials Science*, vol. 68, pp. 314–319, 2013.

[77] R. W. Grosse-Kunstleve, N. K. Sauter, N. W. Moriarty, and P. D. Adams, "The *Computational Crystallography Toolbox*: crystallographic algorithms

in a reusable software framework," in *Journal of Applied Crystallography*, vol. 35, no. 1, pp. 126–136, 2002.

[78] R. Dinnebier and P. Scardi, "X-ray powder diffraction in education. Part I. Bragg peak profiles," in *Journal of Applied Crystallography*, vol. 54, no. 6, pp. 1811–1831, 2021.

[79] V. V. Prasolov, "Polynomials of a Particular Form," in *Polynomials*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 77–131.

[80] J. Oppliger, M. M. Denner, J. Küspert, R. Frison, Q. Wang, A. Morawietz, O. Ivashko, A.-C. Dippel, M. von Zimmermann, N. B. Christensen *et al.*, "Weak-signal extraction enabled by deep-neural-network denoising of diffraction data," 2022, arXiv:2209.09247. URL: https://arxiv.org/abs/2209.09247 last retrieved 2023-09-12.

[81] J. Schuetzke, N. J. Szymanski, and M. Reischl, "A Critical Review of Neural Networks for the Identification of Powder X-ray Diffraction Patterns," in *Advances in X-ray Analysis, Proceedings of the Denver X-ray Conference*, Rockville, MD, USA, 2022.

[82] J. Schuetzke, N. J. Szymanski, G. Ceder, and M. Reischl, "A Critical Review of Neural Networks for the Use with Spectroscopic Data," in *European Crystallographic Meeting (ECM)*, Versailles, France, 2022, DOI: https://dx.doi.org/10.5445/IR/1000151442.

[83] J. Schuetzke, N. J. Szymanski, and M. Reischl, "Validating neural networks for spectroscopic classification on a universal synthetic dataset," in *npj Computational Materials*, vol. 9, no. 1, p. 100, 2023.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[85] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations*, San Diego, CA, USA, 2015.

[86] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 1–9.

[87] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” in *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[88] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, “TensorFlow: Large-scale machine learning on heterogeneous distributed systems,” 2016, arXiv:1603.04467. URL: https://arxiv.org/abs/1603.04467 last retrieved 2023-12-20.

[89] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, “PyTorch: An imperative style, high-performance deep learning library,” 2019, arXiv:1912.01703. URL: https://arxiv.org/abs/1912.01703 last retrieved 2023-12-20.

[90] G. Brauer and H. Gradinger, “Über heterotype mischphasen bei seltenerdoxyden. i,” in *Zeitschrift für anorganische und allgemeine Chemie*, vol. 276, no. 5, pp. 209–226, 1954.

[91] M. Yashima and S. Kobayashi, “Positional disorder of oxygen ions in ceria at high temperatures,” in *Applied Physics Letters*, vol. 84, no. 4, pp. 526–528, 2004.

[92] R. J. Clément, Z. Lun, and G. Ceder, “Cation-disordered rocksalt transition metal oxides and oxyfluorides for high energy lithium-ion cathodes,” in *Energy & Environmental Science*, vol. 13, no. 2, pp. 345–373, 2020.

[93] A. Urban, J. Lee, and G. Ceder, "The Configurational Space of Rocksalt-Type Oxides for High-Capacity Lithium Battery Electrodes," in *Advanced Energy Materials*, vol. 4, no. 13, p. 1400478, 2014.

[94] J. Lee, D. A. Kitchaev, D.-H. Kwon, C.-W. Lee, J. K. Papp, Y.-S. Liu, Z. Lun, R. J. Clément, T. Shi, B. D. McCloskey *et al.*, "Reversible Mn2+/Mn4+ double redox in lithium-excess cathode materials," in *Nature*, vol. 556, no. 7700, pp. 185–190, 2018.

[95] B. Ouyang, N. Artrith, Z. Lun, Z. Jadidi, D. A. Kitchaev, H. Ji, A. Urban, and G. Ceder, "Effect of fluorination on lithium transport and short-range order in disordered-rocksalt-type lithium-ion battery cathodes," in *Advanced Energy Materials*, vol. 10, no. 10, p. 1903240, 2020.

[96] L. C. Pathak and S. K. Mishra, "A review on the synthesis of y–ba–cu-oxide powder," in *Superconductor Science and Technology*, vol. 18, no. 9, pp. R67–R89, 2005.

[97] N. J. Szymanski, P. Nevatia, C. J. Bartel, Y. Zeng, and G. Ceder, "Autonomous and dynamic precursor selection for solid-state materials synthesis," in *Nature communications*, vol. 14, no. 1, p. 6956, 2023.

[98] V. Stanev, V. V. Vesselinov, A. G. Kusne, G. Antoszewski, I. Takeuchi, and B. S. Alexandrov, "Unsupervised phase mapping of x-ray diffraction data by nonnegative matrix factorization integrated with custom clustering," in *npj Computational Materials*, vol. 4, pp. 1–10, 2018.

[99] J. R. Hattrick-Simpers, J. M. Gregoire, and A. G. Kusne, "Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge," in *APL Materials*, vol. 4, no. 5, p. 053211, 2016.

[100] L. A. Baumes, M. Moliner, and A. Corma, "Design of a full-profile-matching solution for high-throughput analysis of multiphase samples through powder x-ray diffraction," in *Chemistry – A European Journal*, vol. 15, no. 17, pp. 4258–4269, 2009.

[101] M. P. Schilling, T. Scherr, F. R. Münke, O. Neumann, M. Schutera, R. Mikut, and M. Reischl, "Automated annotator variability inspection for biomedical image segmentation," in *IEEE Access*, vol. 10, pp. 2753–2765, 2022.

[102] L. Rettenberger, N. J. Szymanski, Y. Zeng, J. Schuetzke, S. Wang, G. Ceder, and M. Reischl, "Uncertainty-aware particle segmentation for electron microscopy at varied length scales," in *npj Computational Materials*, vol. 10, no. 1, p. 124, 2024.

[103] B. H. Toby and R. B. Von Dreele, "*GSAS-II* : the genesis of a modern open-source all purpose crystallography software package," in *Journal of Applied Crystallography*, vol. 46, no. 2, pp. 544–549, 2013.

[104] H. M. Rietveld, "A profile refinement method for nuclear and magnetic structures," in *Journal of Applied Crystallography*, vol. 2, no. 2, pp. 65–71, 1969.

[105] E. Grossi and M. Buscema, "Introduction to artificial neural networks," in *European journal of gastroenterology & hepatology*, vol. 19, no. 12, pp. 1046–1054, 2007.

[106] C. Cortes and V. Vapnik, "Support-vector networks," in *Machine learning*, vol. 20, pp. 273–297, 1995.

[107] L. Breiman, "Random forests," in *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[108] F. R. Münke, J. Schuetzke, F. Berens, and M. Reischl, "A Review of Adaptable Conventional Image Processing Pipelines and Deep Learning on limited Datasets," in *Machine Vision and Applications*, vol. 35, no. 2, p. 25, 2024.

[109] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.

[110] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," in *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, 2020.

[111] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1251–1258.

[112] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 618–626.

[113] S. Sheshachala, B. Huber, J. Schuetzke, R. Mikut, T. Scharnweber, C. M. Domínguez, H. Mutlu, and C. M. Niemeyer, "Charge controlled interactions between dna-modified silica nanoparticles and fluorosurfactants in microfluidic water-in-oil droplets," in *Nanoscale Advances*, vol. 5, no. 15, pp. 3914–3923, 2023.

[114] T. Phan-Xuan, S. Schweidler, S. Hirte, M. Schüller, L. Lin, A. D. Khandelwal, K. Wang, J. Schuetzke, M. Reischl, C. Kübele *et al.*, "Using the high-entropy approach to obtain multimetal oxide nanozymes: Library synthesis, in silico structure–activity, and immunoassay performance," in *ACS Nano*, vol. 18, no. 29, pp. 19 024–19 037, 2024.

[115] J. Schuetzke, A. Benedix, R. Mikut, and M. Reischl, "Siamese Networks for 1D Signal Identification," in *Proceedings 30. Workshop Computational Intelligence, Berlin, Germany, November 26-27, 2020.* Karlsruhe, Germany: KIT Scientific Publishing, 2020, pp. 17–31.

[116] J. Schuetzke, B. Jones, N. Henderson, N. Rodesney, A. Benedix, K. Knorr, R. Mikut, and M. Reischl, "Application of Machine Learning to XRD Phase Identification," in *Denver X-Ray Conference*, Denver, CO, USA, 2020, DOI: https://dx.doi.org/10.5445/IR/1000127718.