crystals

MDPI

Article

Learning Self-Supervised Representations of Powder-Diffraction Patterns

Shubhayu Das , Markus Vorholt , Andreas Houben and Richard Dronskowski *

Institute of Inorganic Chemistry, RWTH Aachen University, D-52056 Aachen, Germany;
shubhayu.das@ac.rwth-aachen.de (S.D.); markus.vorholt@ac.rwth-aachen.de (M.V.);
andreas.houben@ac.rwth-aachen.de (A.H.)
* Correspondence: drons@HAL9000.ac.rwth-aachen.de

Abstract: The potential of machine learning (ML) models for predicting crystallographic symmetry information from single-phase powder X-ray diffraction (XRD) patterns is investigated. Given the scarcity of large, labeled experimental datasets, we train our models using simulated XRD patterns generated from crystallographic databases. A key challenge in developing reliable diffraction-based structure-solution tools lies in the limited availability of training data and the presence of natural adversarial examples, which hinder model generalization. To address these issues, we explore multiple training pipelines and testing strategies, including evaluations on experimental XRD data. We introduce a contrastive representation learning approach that significantly outperforms previous supervised learning models in terms of robustness and generalizability, demonstrating improved invariance to experimental effects. These results highlight the potential of self-supervised learning in advancing ML-driven crystallographic analysis.

Keywords: diffraction; crystallography; machine learning; self-supervised learning; representation learning

check for updates

Academic Editor: David Cordes

Received: 21 February 2025
Revised: 15 April 2025
Accepted: 21 April 2025
Published: 23 April 2025

Citation: Das, S.; Vorholt, M.; Houben, A.; Dronskowski, R. Learning Self-Supervised Representations of Powder-Diffraction Patterns. *Crystals* **2025**, *15*, 393. https://doi.org/10.3390/cryst15050393

1. Introduction

Determining an unknown crystal structure and, hence, identifying a new chemical compound (usually called a *crystallographic phase*) from X-ray or neutron powder diffraction data is an inverse problem that requires experienced users to make several strategic choices across multiple steps in the course of structure determination. The typical pipeline following the measurement of intensities from an appropriate laboratory or high-resolution X-ray (or neutron) diffractometer and radiation source starts with binning the diffraction pattern, describing the background, and then identifying the so-called Bragg peak positions from elastic coherent scattering. This is followed by indexing the pattern to identify the crystal system, the unit cell, and then narrowing down the number of potential space groups, based on which a structure-solving algorithm suggests candidate structures, eventually improved or rejected by (Rietveld) structure refinement [1]. Depending on the chemical nature of the sample, the candidate structures are usually determined using approaches going under the names Direct Methods or Patterson Method [2], or other methods such as Simulated Annealing [3], Genetic Algorithms [4], Charge Flipping [5], etc. By doing so, the generally complex structure factors are found, which reconstruct (upon Fourier transformation) the measured intensities, so the model may be compared to the real world that has been measured.

The aforementioned *pipeline* carried out by human beings has allowed for a huge number of structure determinations (by solving the *phase problem*, that is, determining the atomic positions in real space), but the exact course is difficult to predict from a more general perspective since it depends on the chemical nature, on the structural complexity, on human skills, and so forth. That being said, specific choices must be made that may be considered as iterative informed guesses, most often also based on additional sources of information. This is particularly true if more than one chemical species (multiple-phase diffraction patterns) is being looked at, so the indexing step faces a tremendous challenge to begin with: which of the Bragg peaks belongs to which phase? It may therefore be a good idea to utilize data-driven machine learning models that can *directly* generate or identify or at least estimate candidate crystallographic structures based on the measured data.

Typically, a supervised machine learning model needs large amounts of labeled data, which, in our specific case, relates to experimental diffraction patterns and the corresponding structure information of the measured sample in order to make accurate predictions. Models trained on limited data can have a strong bias and, therefore, generalize poorly. It is difficult to satisfactorily quantify the amount of data needed, as often even models trained on a seemingly large amount of data can be vulnerable to *natural adversarial examples* [6], i.e., naturally occurring instances that unintentionally cause a model to make incorrect predictions. It may be argued that such models make predictions based on spurious correlations in the data rather than useful features. While limited data availability is a persistent challenge in many experimental sciences, machine-learning models have demonstrated remarkable flexibility in adapting to these constraints. Despite initial hurdles, models like AlphaFold [7] have revolutionized their respective fields by leveraging innovative training strategies and structural priors, ultimately surpassing traditional approaches. Inspired by such advancements, we explore how data-driven models can be designed to tackle the complexities of crystallographic structure prediction.

From a statistical point of view, acquiring *experimental* data representative enough to model the joint distribution of the diffraction patterns (including instrumental effects) and the corresponding structure information is expensive because one cannot quickly synthesize and re-measure hundreds of thousands of solid-state chemical compounds. We can, however, extract experimental crystal-structure information already stored as Crystallographic Information Files (CIF), being part of large-scale crystal-structure databases such as the Crystallography Open Database (COD) or the Inorganic Crystal Structure Database (ICSD). Given that the data contained in those CIF files are accurate and idealized (that is, almost free of any measuring inaccuracies), experimental diffraction patterns may be straightforwardly *simulated*. It is important to note, however, that designing a machine-learning model trained with simulated data may introduce biases that make it perform poorly on experimental data.

This issue can, at least to a certain extent, be circumvented by using models that take manual or *handcrafted* features of the experimental diffraction patterns as the input. We shall later show that for specific tasks such as indexing and crystal-system determination such models can potentially outperform well-known existing search-based, non-data-driven algorithms like NTREOR [8]. However, models trained on handcrafted features (typically chosen by a trained human being) are time-consuming and difficult to evaluate since such models require some *pre-processing* steps (e.g., peak detection), making it difficult to debug mistakes. In other words, this real-world component makes the model vulnerable to adversarial effects inherently present in *any* experimental measurement. More importantly, such models and the features used by them do not generalize over different tasks, and therefore cannot be used to build an end-to-end structure prediction model.

To fulfill the ultimate vision of reliably using data-driven ML models for structure solutions from powder diffraction patterns, we need to ensure that the underlying architecture is both scalable and capable of generalizing across different tasks. Therefore, our goal is to use the *entire* diffraction pattern as input and design a model that inherently learns robust feature representations. To achieve this, the model must be invariant to variations in the input caused by sample or instrumental effects and noise while remaining sensitive to variations arising from structural differences. Simply put, changes in the input should only influence the prediction if they correspond to a genuine structural difference in the measured crystallographic structure. For this, we investigate neural network architectures with training pipelines inspired by recent advancements in semi- and self-supervised representation learning. We believe this approach will facilitate previously unseen robustness against *natural adversarial examples* arising from noise to experimental variations in the diffraction pattern input, and one needs to ensure that both the model architecture and the training methodology are suitable to reach this goal.

To focus and streamline our work, we limit our scope to prediction tasks, specifically classifying the crystal system, extinction group, and space group from diffraction patterns of predominantly single-phase inorganic crystal structures. We explore model architectures, testing standards, and training methodologies to enhance model reliability and generalizability.

2. Previous Works

In this Section, we highlight the growing intersection between machine learning and classical crystallographic techniques, motivating further research into integrating modern ML frameworks into crystallographic workflows. Gasparotto et al. [9] introduced the TORO Indexer, a PyTorch-based (version 2.0) indexing algorithm designed for kilohertz serial crystallography. This work leverages computational optimizations commonly found in deep learning frameworks to accelerate the indexing step, making it feasible for high-throughput serial crystallography experiments. Recent works have explored a range of data-driven methodologies, from traditional machine-learning techniques to advanced neural networks, to tackle these challenges. For example, Suzuki et al. [10] studied the use of machine-learning models such as Support Vector Machines, Random Forests, and Randomized Decision Trees trained with handcrafted features such as the position of the first ten low-angle (2θ) peaks and the number of peaks in the 2θ range of 0–90°. The data for training and testing were generated from *simulated* powder-diffraction patterns stemming from CIF files taken from ICSD, using the aforementioned handcrafted features. As is typical for data-driven models, the test set was a part of the entire dataset kept separate from the data used for training the model. Suzuki et al. [10] reported a crystal-system classification accuracy of 90% and a space-group classification accuracy of about 80.5%. The performance reported over *experimental* data, however, was limited to two rather typical laboratory XRD measurements of $Ca_{1.5}Ba_{0.5}Si_5N_6O_3$ and $BaAlSi_4O_3N_5{:}Eu^{2+}$, but space-group classification failed for both. We adopt this approach of using handcrafted features in one of our baseline experiments.

Park et al. [11] proposed treating the entire powder diffraction pattern as a one-dimensional picture using a Convolutional Neural Network (CNN) trained on single-phase simulated diffraction patterns of crystal structures from the ICSD database. This was intended to predict the crystal system, the extinction group (derived from systematic absences or extinctions), and the space group. The authors proposed a data-generation pipeline using a set of fixed parameters such as the structure factor, the multiplicity, the Lorentz polarization factor, and a set of randomly selected parameters involved in the pseudo-Voigt peak profile function, the Caglioti parameters [12], and the coefficients of the background polynomial. Owing to this, the proposed CNN internally learned a representation of the diffraction pattern as opposed to the models presented by Suzuki et al. [10] and their handcrafted features. When tested on generated data using the aforementioned data-generation pipeline on a subset of ICSD crystal structures not used in training, the paper reported 94, 83.8, and 81.1% accuracy for crystal-system, extinction-group, and space-group classification, respectively. For testing on experimental data, Park et al. [11] also reported their model's prediction on $Ca_{1.5}Ba_{0.5}Si_5N_6O_3$ and $BaAlSi_4O_3N_5$:Eu^{2+}. Similar to Suzuki et al. [10], the predictions also failed for both extinction and space groups. Although the model by Park et al. [11] learned better representations due to random parameters of the data-generation pipeline, the authors did not explicitly investigate the model's invariance as regards experimental effects. In addition, the rather limited experimental testing makes it difficult to analyze the model's robustness for practical use.

Lee et al. [13] expanded on some of the ideas by Park et al. [11], first by including training Fully Convolutional Neural Networks (FCNNs) as well as a Vision Transformer-inspired neural network for predicting extinction and space group for single-phase simulated diffraction patterns from the ICSD and the Materials Project (MP) dataset and, second, neural network regression models trained on the non-experimental MP data to estimate the DFT-calculated band gap, formation energy, and the energy above the convex hull. Although the authors reported close to state-of-the-art (SOTA) performances, they also did not perform extensive tests for the model's robustness over natural adversarial examples.

Oviedo et al. [14] proposed a more representative simulation pipeline that models the evolution of a thin-film XRD experiment by modifying the XRD data in terms of pattern shifting, peak scaling, and peak elimination, an approach usually dubbed as *augmentation* in this simulative context. They measured 85 thin-film experimental XRD patterns of known structures belonging to seven different space groups, including perovskite-like materials such as lead halides ($Pm\bar{3}m$), tin halides ($I4/mcm$), Cs-Ag-Bi bromide double perovskites ($Fm\bar{3}m$), and Bi and Sb halides ($P\bar{3}m1$, Pc, $P2_1/c$, $P6_3/mmc$). The goal of Oviedo et al. [14] was to classify the XRD patterns into the seven aforementioned space groups, and the data-augmentation strategy was designed such that the simulated data used for training accurately fitted the measured thin-film X-ray diffraction (XRD) patterns. The paper reported an accuracy of about 99 and 80% on a simulated/experimental test set for the seven space-group classification problem; admittedly, this corresponds to a much simpler task compared to other works.

Salgado et al. [15] trained a neural network model for classifying the space group using a combination of simulated and experimental data. The simulated XRD data, consisting of 1.2 million training patterns, were generated from ICSD structures using a set of Caglioti parameters and noise implementations, and the experimental XRD patterns (908 patterns) were collected from the RRUFF dataset. The authors evaluated space-group classification accuracy across models trained on various sets of data. Particularly, they trained neural network models on simulated data as well as a combination of simulated data and half of the experimental data, while using the rest for testing. The best-case accuracy reported for their models trained entirely using simulated data was 66%, which increased to 77% when half the experimental data were added for training. Hence, Salgado et al. [15] addressed the challenges involved when training machine learning models on simulated data that can be robust enough to work on experimental data.

Lolla et al. [16] introduced a semi-supervised deep learning model for classifying powder neutron diffraction patterns into 14 Bravais lattices and 144 space groups. The model leveraged simulated diffraction patterns as labeled data while exploring the use of partially labeled datasets during training. Despite achieving state-of-the-art results on simulated test datasets, the study did not incorporate real experimental data as labeled training examples, nor did it evaluate the model's performance on real experimental datasets. This limitation highlights a gap in generalizability to real-world XRD patterns, which may include experimental noise and other complexities not captured in simulations. However, this work introduced an interesting idea of using the Discriminator of a Generative Adversarial Network [17] to predict the Bravais lattice and space groups. This concept introduces the possibility of incorporating experimental XRD patterns into the unsupervised training modes of the Discriminator in future work, thereby potentially improving its applicability to real-world data.

More recently, Siamese networks [18,19] and contrastive learning [18] have emerged as promising techniques for learning robust representations of crystallographic data. These approaches aim to learn similarity-based embeddings, making them particularly well-suited for tasks where manually labeled experimental data are sparse or expensive. Inspired by these ideas, we investigate contrastive learning methods such as SimCLR and Barlow Twins, which encourage models to learn representations invariant to experimental noise while preserving structural information.

Furthermore, recent advancements in end-to-end crystal structure generation [20] demonstrate the feasibility of using ML models to directly predict structures from powder X-ray diffraction patterns. These approaches highlight the potential of fully data-driven methods, moving beyond classification tasks toward complete structure determination. Lai et al. (2025) [20] use a contrastive learning approach to align representations of powder XRD patterns and their corresponding crystal structure representations while using these representations for conditioning a diffusion-based [21,22] crystal structure generative model.

Our work focuses more specifically on learning invariant representations of powder XRD patterns. More specifically, in this paper, we adopt a self-supervised representation learning strategy, which relies entirely on simulated data for training.

3. Crystallographic and Diffraction Data

The data-driven models discussed in this paper are trained and tested using data simulated from known crystallographic structures. It is important to outline the details of the simulation process to provide context for the subsequent analysis. The general functional form of the calculated intensity y_{ci} at the i^{th} position of the diffraction pattern for a given crystal structure is given as:

$$\lambda = 2d \cdot \sin\theta \tag{1}$$

$$y_{ci} = y_{bi} + s \sum_{k\,\epsilon\,\{hkl\}} LP_k \cdot M_k \cdot |F_k|^2 \cdot \Phi(2\theta_i - 2\theta_k) \cdot A \cdot E \tag{2}$$

$$F_k = \sum_{j\,\epsilon\,\{\text{atoms}\}} N_j \cdot f_j \cdot \exp\left[2\pi i\left(k \cdot \vec{x}_j\right)\right] \cdot \exp\left[-B(\sin\theta/\lambda)^2\right] \tag{3}$$

Here, θ is the scattering angle, λ is the wavelength, d is the interplanar spacing, s is the scale factor, k is the reciprocal lattice vector as expressed by the hkl triple, LP_k are the Lorentz and polarization factors, M_k the multiplicity factor, F_k is the structure factor, Φ is the profile function, and $A \cdot E$ is the product of the absorption and extinction factors. The structure factor contains all the crystallographic information about the atoms j and their relative positions in the unit cell, while the set of all k dictates the symmetry of the unit cell. Here, N_j is the multiplicity of the atomic position, f_j is the atomic form factor, $\vec{x}_j$ is the position vector, and B is the (isotropic) thermal displacement factor, also going under the name Debye–Waller factor. We use the Inorganic Crystal Structure Database (ICSD), which consists of over 200,000 experimentally determined and carefully curated Crystallographic Information Files (CIFs). Based on the latter, we simulate their XRD patterns using a fully automated pipeline. We first pre-compute only the structure factor contribution to the calculated integrated intensities for all ICSD CIFs, using the scriptable interface of GSAS-II [23,24]. This simply gives us some discrete intensities at specific 2θ positions or d spacings. The diffraction data used in the models discussed in this paper all work with these precomputed integrated intensities.

Models utilizing the full profile of the diffraction pattern as input individually compute the diffraction patterns for both training and robust testing. Each component of the profile—such as the profile function, peak widths, various sample effects, and noise—is treated independently either as signal-processing operations or as *augmentations*, each governed by distinct parameters. The specific details of these operations will be outlined as we describe each computational experiment in Section 4.

Generally, all data used for training and testing are kept separate to ensure unbiased evaluation. The model's performance and robustness are assessed across naturally occurring adversarial examples. To conceptualize this, we imagine diffraction patterns (the data) to exist in a virtual two-dimensional plane as defined by two orthogonal axes, one that affects the pattern due to structural changes and another one that affects it due to experimental or sample effects. Moving along one axis corresponds to traversing through the diffraction patterns of all feasible crystal structures, while moving along the other axis represents *variations* in the diffraction patterns for a single crystal structure.

We refer to the latter as the *invariance* and the former as the *equivariance* axes. A model is *invariant* to a transformation if its output remains unchanged when the input undergoes that transformation (e.g., a classifier that recognizes a crystal system regardless of rotation). A model is *equivariant* to a transformation if its output changes in a predictable way when the input undergoes that transformation (e.g., a network that shifts peak positions in an XRD pattern in response to a change in lattice parameters). The model's goal is to extract feature representations that are equivariant to structural differences in the input (the XRD pattern) but invariant to variations caused by noise, sample, or experimental effects.

In addition to the large number of simulated XRD patterns, we also collected 82 experimentally measured XRD diffraction patterns from semi-pure chemical samples found in our own laboratories, using STADI-P and STADI-MP powder diffractometers (STOE and Cie GmbH, Darmstadt, Germany) equipped with Cu-Kα_1 and Mo-Kα_1 radiation, respectively. The data cover all seven crystal systems and patterns with varied levels of noise. Out of this entire set, we consider 18 XRD patterns to be significantly noisy, real-life datasets in their original meaning. Unlike the simulated data, where each aspect of the diffraction pattern follows a precise functional form, the experimental measurements exhibit functional variability across different aspects of the full profile. These 82 experimentally measured XRD patterns are reserved exclusively for evaluation and testing purposes, not for training.

4. Computational Experiments

This Section presents a series of computational experiments aimed at predicting symmetry information and lattice parameters from diffraction patterns. The objective is to explore the inherent complexity of the problem and evaluate different approaches to identify the most robust and principled method. Figure 1 provides an overview of the entire setup and its respective methodologies.

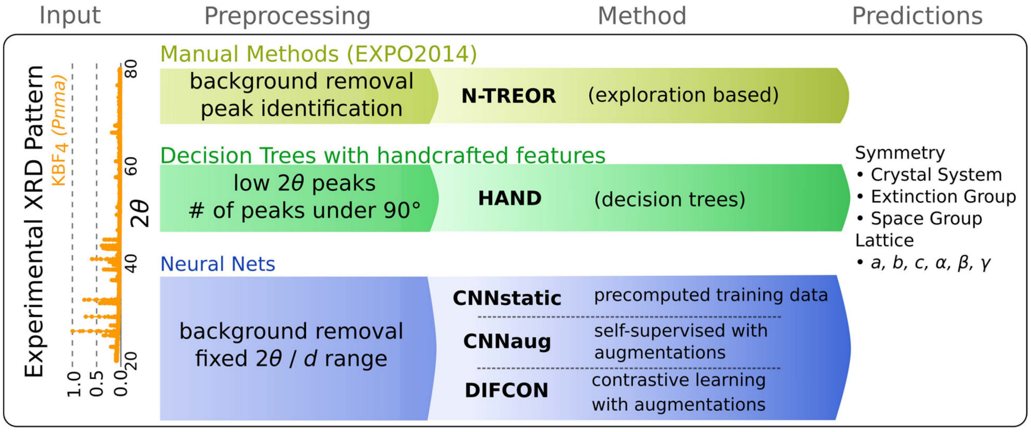

Figure 1. Overview of the computational methodologies showing the input, preprocessing, principal method, and predictions.

4.1. NTREOR

We begin by using the NTREOR indexing algorithm [8] as used in the EXPO2014 software package (version 1.22.11), an updated version of EXPO2013 [25]. NTREOR is a heuristic, well-known, highly tested, and trustworthy optimization method employing a trial-and-error approach to search for solutions in the index space by varying the Miller indices. It requires the user to accurately identify the peak positions in the diffraction pattern, which are then used to predict the crystal class, the extinction symbol, and its lattice parameters. NTREOR can propose multiple candidate solutions, each accompanied by the cell volume and a figure of merit to help identify the most likely solution. In particular, we use M_{20}, the de Wolff figure of merit, to evaluate the quality of each candidate in the following.

We apply this method to index the 82 experimental diffraction patterns detailed in Section 3. For each pattern, the radiation wavelength is specified, and peak positions are manually identified. The algorithm is then executed, and the best candidate solution is selected based on the highest figure of merit. In some cases where multiple solutions have the same figure of merit, we pick the solution with the smallest cell volume. This approach serves as a baseline for comparison with the data-driven methods that will be introduced in the subsequent Sections.

As regards the statistical interpretation of the results of the NTREOR algorithm applied to those manually identified reflections, NTREOR achieves an accuracy of 49% in predicting the crystal system, correctly classifying 40 out of 82 XRD patterns. Only for these 40 correctly classified patterns, we calculate the root-mean-squared error (RMSE) of the lattice parameters. The mean RMSE for the cell vectors is 1.38 Å, with a standard deviation (σ) of 2.46 Å, while the mean RMSE for the cell angles is 0.81°, with σ = 1.507°; 22 of the 40 correctly classified patterns exhibit an RMSE of the cell vectors below 0.1 Å, and 24 of the 40 patterns have an RMSE of the cell angles below 0.1°. Notably, this rather simple strategy does not cover those cases where the NTREOR solution corresponds to one out of many supercells of the correct unit cell. Likewise, it will not count an almost correct sub-cell of the correct unit cell as a successful case.

4.2. HAND

The first data-driven model we train, HAND, is inspired by the methods proposed by Suzuki et al. [10]. This model uses handcrafted selected features of the diffraction patterns as inputs, specifically the first ten peak positions in the low 2θ range and the total number of peaks within the 2θ = 0–90° range for constant copper radiation at $\lambda_{K\alpha}$ = 1.5418 Å, in compliance with the copper wavelength used by ICSD. As mentioned in Section 3, we use GSASIIscriptable [24] for this. Compared to models that process the entire diffraction pattern, HAND is less complex, employing a simpler architecture based on Randomized Decision Trees. The model is designed to predict the crystal system (CS), the extinction group (EG), and the space group (SG) separately. The crystallographic data from the ICSD are randomly split, with 90% used for training and the remaining 10% reserved for testing. By testing on a distinct set of crystal structures, we aim to evaluate the model's equivariance to variations caused by changes in crystal structure. Our most-developed, i.e., best-fitted, model achieves an accuracy of 94% for classifying the crystal system, 91% for classifying the extinction group, and 87% for classifying the space group. We notice that in 3.5% of the test instances, the predicted space group did not belong to the correctly predicted crystal system. This highlights the contribution of spurious correlations in the data (the handcrafted features) toward the model's prediction. Notably, the model performance does not drop significantly with perturbations to the peak position, suggesting a better degree of robustness compared to NTREOR. When testing the model on the experimental test set and using the manually identified peaks as in the previous experiment with NTREOR, we observe an accuracy of 55.5%, 32%, and 39.5% for the crystal system, extinction group, and space group prediction tasks, respectively. The crystal system classification accuracy shows an improvement over that of the NTREOR on the same experimental test set. It is worth noting, however, that the space-group classification accuracy is *larger* than that of the extinction-group classification. Further investigation shows that in 26% of the experimental test instances, the predicted space group did not belong to the correctly predicted crystal system. This further validates our claim that the model relies on spurious correlations.

4.3. CNN Static-Supervised

As motivated in the introduction, our main goal is to design models that require minimum manual/human involvement and can be scaled in the future for more complicated structure solution tasks. To that end, in this computational experiment, we aim to address the issues when utilizing the full profile of the diffraction pattern. The data are generated using simulated Bragg reflections (as described in Section 3) while incorporating practical experimental and sample effects as signal processing operations. These operations are carefully designed to capture the typical characteristics of *noise* and the measurement specifics of a standard laboratory constant-wavelength X-ray diffractometer. We emphasize that this approach is a crucial aspect of our work and will be applied in subsequent experiments, as we explore the feasibility of training our models entirely on simulated XRD data.

To begin with, we employ a convolutional neural network (CNN) to process the input diffraction pattern, treating it as a one-dimensional image. The architecture is based on ResNet [26] by He et al., a popular deep-learning framework designed to mitigate the vanishing gradient problem in deep neural networks. The model incorporates multiple residual blocks, which feature skip connections to facilitate the flow of gradients during backpropagation. These skip connections allow the network to learn residual mappings instead of direct mappings, improving convergence and enabling the training of deeper architectures.

Furthermore, we design the model to simultaneously predict the crystal system (CS), extinction group (EG), and space group (SG). The model employs a multi-prediction head architecture: the CS head predicts the crystal system (7 classes), the EG head predicts the extinction group (101 classes), and the SG head predicts the space group (230 classes in total), utilizing the output from the CS head's prediction. The mapping from CS to SG is injective and follows the crystallographic structure, that is, *triclinic*: SG $\in$ [1–2], *monoclinic*: SG $\in$ [3–15], *orthorhombic*: SG $\in$ [16–74], *tetragonal*: SG $\in$ [75–142], *rhombohedral/trigonal*: SG $\in$ [143–167], *hexagonal*: SG $\in$ [168–194], and *cubic*: SG $\in$ [195–230].

In total, the model features 8 prediction heads, of which 7 are for the CS and SG, and 1 for the EG. This design introduces an *inductive bias*, aligning the model's architecture and its inherent structure of the task, as commonly discussed in machine learning literature. The output of each classification head $\left(x_{cs}, x_{eg}, x_{sg}\right)$ is finally passed through a so-called softmax function, to output a set of probabilities (p_{CS}, p_{EG}, p_{SG}) over the respective number of classes N.

$$\mathrm{softmax}(x) = e^{x} / \sum_{i=1}^{N} e^{x} \tag{4}$$

The loss function at each classification head uses a cross-entropy (CE) loss, which compares the predicted probabilities p to the true labels $\hat{p}$.

$$\mathrm{CE} \ = -\frac{1}{N}\sum\nolimits_{i=1}^{N} \hat{p}_i \cdot \ ln\ p_i \tag{5}$$

The final loss function is a weighted average of $\left(\mathrm{CE}_{CS}, \mathrm{CE}_{EG}, \mathrm{CE}_{SG}\right)$. The weights are treated as hyperparameters and tuned independently of each other during training.

The XRD data used for training and testing are generated through a simulation pipeline beginning with pre-computed ideal Bragg peak positions and calculated intensities $\{2\theta_i, y_{ci}\}$ for a copper radiation wavelength of $\lambda_{K\alpha}$ = 1.5418 Å; similar to the previous experiment, this is performed using GSASIIscriptable [24]. These values are then utilized to construct the full profile XRD pattern through a series of parameterized signal processing operations, which we will now outline. Most of these operations are inspired by instrument effects, while some are intended to facilitate and validate a more thorough generalizability.

- We begin with a simple *zero shift*, shifting the entire pattern along the 2θ axis by a small (maximally ±0.02°) amount, simulating errors commonly introduced by improper calibration of the instrument.
- Next, we add a very small amount of random-like *x-axis white noise* to each peak position ($2\theta_i$).
- *Peak cropping and padding*: the peaks at the tails (high and low 2θ) are randomly cropped; i.e., the y_{ci} are replaced with zeros.
- *Noise to the integrated intensities* is also introduced, simulating all sorts of effects, e.g., of non-ideal detectors. Operations that simultaneously vary the position and integrated intensity effectively model variations due to different experimental effects. For example, changes in integrated intensities can at least partly represent the impact of micro-strain anisotropies, preferred orientations in the powder sample, wavelength fluctuations, as well as Lorentz and polarization factors.
- *Binning* is performed over the 2θ = 0–120° range using a fixed bin width and a copper radiation wavelength of $\lambda_{K\alpha} = 1.5418$ Å. While this might seem trivial, it requires careful consideration. The peak positions depend on the radiation wavelength via Bragg's law, and the choice of binning affects intensity values. To enable the model to generalize across different radiation wavelengths, one could use the d-spacing for peak positions. Due to the non-linear relationship between d-spacing and 2θ, however, variations in the binned intensities become highly pronounced and cannot be sufficiently accounted for by adding noise to the integrated intensities before binning. Our investigation shows that the variation over the binned intensities is more reasonable when binning over 2θ for a specific λ being equivalent to the wavelength used for training. For inputs with a different radiation wavelength λ', the measured $2\theta'$ value can be easily converted to the corresponding 2θ of the training wavelength λ via the d-spacing, namely, $2\theta = 2\text{arcsin}\left[\frac{\lambda}{\lambda'}\sin\left(\frac{2\theta'}{2}\right)\right]$. Please note that this might truncate high $2\theta'$ values for the case $\lambda' < \lambda$.
- *Small impurity peaks*: small amounts of random-like impurity peaks whose intensities are smaller or comparable to the smallest Bragg reflection, but higher than the noise level in a diffraction pattern, are added. This acts as a type of compositional noise in the XRD profile.
- *Convolving a peak profile* (Ω): a *pseudo-Voigt* profile with peak asymmetry is convolved across the binned diffraction pattern. This is inspired by the profile used by the CW-XRD refinement program in GSAS-II [23,24], although here we are simply interested in a function form that offers reasonable variations of the profile and not its precise fitting capabilities. The full-width-at-half-maximum (FWHM) is inspired by the Caglioti [12] functional form and is presented in Table 1; here, the parameters $U, V, W,$ and p are sampled using Latin-Hypercube sampling [27]; the *pseudo-Voigt* profile is a linear combination of a Gaussian and a Lorentzian using the same (FWHM) weighted by the mixing parameter η; asymmetry is introduced by an error function applied over the *pseudo-Voigt* profile.

Table 1. Overview of the signal processing operations used for simulating XRD patterns. Here, $\mathcal{U}(\;)$ signifies uniform distribution within a specific range and $\mathbb{R}$ defines real numbers of a specific dimension. The units of the variables refer to the physical parameters and have been dropped for reasons of simplicity.

Operation	Functional Form/Random Variable(s)
zero shift	$\delta_{2\theta} \sim \mathcal{U}(-0.02, 0.02)$
x-axis white noise	$2\theta_{\text{new}} = 2\theta_{\text{old}} + \delta_{2\theta}$, $\delta_{2\theta} \in \mathbb{R}^{\text{no. of peaks}}$, $\delta_{2\theta} \sim \mathcal{U}(-a, a)$, $a \sim \mathcal{U}(-0.005, 0.005)$
peak cropping and padding	$y_i = 0$, if $i > 2\theta_{high}$ and $i < 2\theta_{\text{low}}$, $2\theta_{\text{low}} \sim \mathcal{U}(0.0, 0.2)$, $2\theta_{\text{high}} \sim \mathcal{U}(100, 120)$
intensity noise	$y_{\text{new}} = y_{\text{old}} + \delta_y \cdot y_{\text{old}}$, $\delta_y \in \mathbb{R}^{\text{no. of peaks}}$, $\delta_y \sim \mathcal{U}(-\ell, \ell), \ell \sim \mathcal{U}(-0.005, 0.005)$
Binning (no random variables)	no. of bins : $N =$ 12,000, index. of each bin : j, $2\theta_j = 0.01 \times (j-1) + 0.005$ $x_{\text{binned}_j} = \begin{cases} 0, & \text{if no Bragg peaks in } j^{th}\text{bin} \\ \text{mean}(y_j), & y_j = \text{Bragg peaks in the } j^{th} \text{ bin} \end{cases}$
impurity peaks	no. of impurity peaks : $M \in \mathbb{R}^1[0, 6]$, $x_{\text{impurity}_j} \in \mathbb{R}^M$, $x_{\text{impurity}_j} \sim \mathcal{U}(0.01, 0.05)$
profile	$\text{FWHM}(U, V, W, p) \approx \sqrt{U \cdot \tan^2\theta + V \cdot \tan\theta + W + p/\cos^2\theta}$ $\text{FWHM} \in \mathbb{R}^1[0.005, 0.05]$, $\text{pV}(\text{FWHM}, \eta) = (1-\eta) \cdot \text{Gaussian} + \eta \cdot \text{Lorentzian}$ $\eta \sim \mathcal{U}(0, 1)$, asymmetry coefficient : $\rho \sim \mathcal{U}(-1, 1)$, $\Omega \sim \rho \cdot \text{sigmoid} \cdot \text{pV}$
background noise	dilation : $d \sim \mathcal{U}(0, 10)$, $x_{\text{white noise}} \in \mathbb{R}^d$, $x_{\text{white noise}} \sim \mathcal{U}(-n, n)$, $n \in \mathcal{U}(0, 0.05)$
cropping and padding	$x_j = 0$, for $2\theta_{\text{high}} < 2\theta_j < 2\theta_{\text{low}}$, $2\theta_{\text{low}} \sim \mathcal{U}(0, 24)$, $2\theta_{\text{high}} \sim \mathcal{U}(100, 120)$

- *Overall noise*: a background noise is added to the XRD profile. This contains a combination of white noise and intensity-dependent noise. The pattern is then re-normalized.
- *Cropping and Padding*: Finally, the edges of the XRD profile are cropped randomly and padded with zeros. This is to facilitate generalizability over cases where the edges of the XRD pattern need to be cropped due to extreme background radiation.

Figure 2 shows the effects of some of these operations on the diffraction pattern. Notably, we consider the aforementioned operations to be parameterized by random variables. In most cases, these random variables are sampled from a uniform distribution with reasonable ranges and are detailed in Table 1. We implement these operations in PyTorch (version 2.0), and the simulation is performed by using configuration files that specify these aforementioned ranges. Naturally, we have separate configuration files for simulating *training* and *testing* data, ensuring no overlap between the parameters sampled for the aforementioned operations. We, therefore, call this the *invariance test set*. We carry out the simulation with 90% of the ICSD crystal structures (≈180,000 structures), along with our non-overlapping parameter sampling strategy to simulate about 1.08 million XRD patterns (six diffraction patterns per ICSD structure) for the *training set* and about 360,000 XRD patterns (two diffraction patterns per ICSD structure) for the *invariance test set*. We also have the usual *equivariance test set* that simply uses XRD patterns simulated from 10% of the ICSD structures (≈20,000 structures) that are kept separate from those used in training. We then simulate about 40,000 XRD patterns (two diffraction patterns per ICSD structure) by sampling using the *training set* simulation parameters. Further implementational details can be found in our code (see Data and Code Availability Statement).

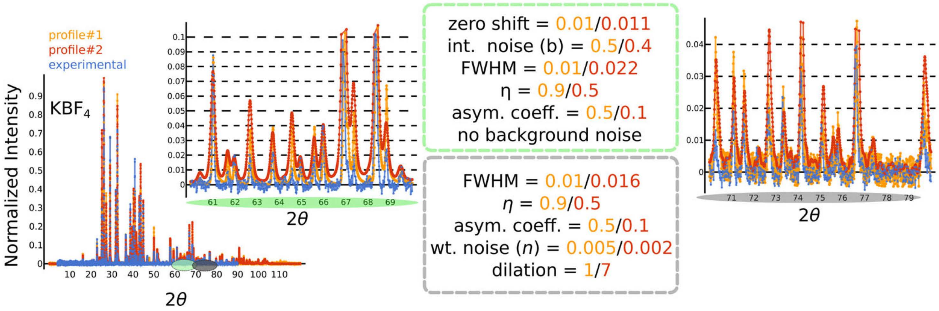

Figure 2. Exemplary comparison of the experimental pattern of KBF$_4$ (blue) with two randomly selected ablations (red, orange) according to the given parameters.

Together, this follows our robust testing strategy discussed earlier in Section 3. The *equivariance test set* is also considered a validation set, which is used to track the training progress and stop training when this accuracy starts to decrease to prevent the model from overfitting.

For the *equivariance test set*, the classification accuracy of the **CNNstatic** model for the crystal system (CS), extinction group (EG), and space group (SG) is 89%, 82%, and 79%, respectively. However, for *the invariance test set*, the classification accuracy drops significantly to 40% for CS, 33% for EG, and 24% for SG. These results indicate that the model struggles to generalize effectively when faced with variations introduced by the aforementioned experimental effects.

Additionally, we also test the model in our *experimental test set*, containing 82 diffraction patterns whose wide variety of background radiation is described manually for each pattern. It is prudent to mention here that even after the background subtraction, there is often a significant amount of background noise. The testing performed here can therefore validate the effectiveness of our simulation pipeline in modeling effects due to significant background noise. However, the classification accuracy for CS, EG, and SG on the experimental test set is only 22%, 15%, and 13%, respectively.

These results demonstrate that despite extensive efforts to model the signal processing details in the simulation pipeline and to train a CNN model accurately, the performance remains unsatisfactory. Any further attempts to tune the model would likely lead to an overfitting to the specific training data distribution.

4.4. CNN with Augmentations

This computational experiment approaches the challenge of designing a data-driven powder XRD indexing algorithm from a novel perspective. To begin, we draw on established ontology in the field of machine learning (ML), particularly concerning the relationship between inputs and outputs in such models.

In traditional ML domains, such as computer vision and natural language processing, models are categorized based on the nature of the dataset:

- **Supervised Learning**: both inputs and outputs are fully labeled for all instances in the dataset.
- **Unsupervised Learning**: outputs are entirely unlabeled, and the model discovers patterns or structures in the data without explicit guidance.
- **Semi-Supervised Learning**: a portion of the data is labeled, while the rest remains unlabeled.
- **Self-Supervised Learning**: partial relationships between inputs and outputs are leveraged during different stages of the training pipeline. Self-supervised models generate pseudo-labels or pretext tasks to aid in learning meaningful representations.

In the context of indexing XRD measurements—or structure determination more broadly—we encounter a unique challenge, namely, the lack of sufficient real experimental data for training a reliable ML model, as alluded to already. This necessitates the use of simulated data derived from known crystal structures. Unlike conventional ML applications, this problem involves a peculiar inversion where the *output* (crystal structure information) is known, but the *input* (the XRD pattern) must be generated or simulated from the output.

Furthermore, certain aspects of the input—such as noise, background variation, and experimental inconsistencies in XRD patterns—cannot be *reliably* simulated from the crystallographic information file alone. This can be pictured using an analogy: training an ML model on synthetically simulated diffraction patterns and expecting it to perform robustly on real experimental data is akin to training a model to distinguish between trees with and without leaves using only simplistic, childlike drawings of trees.

This analogy underscores the inherent challenges and complexity of the task at hand, driving the exploration of self-supervised learning techniques to bridge the gap between synthetic training data and real-world experimental data. In this Section, we start to address this gap by designing a training pipeline that incorporates self-supervised learning principles to enhance the model's robustness and generalizability.

We build upon the model presented in the previous experiment Section (**CNNstatic**) by making targeted modifications while retaining the core model architecture, which features multiple classification heads and employs the same loss function. The primary change lies in how the data-generation pipeline is integrated into the training process.

While the Bragg peaks (both positions and integrated intensities) are still preprocessed before training, the signal-processing operations generating the full-profile diffraction patterns are now incorporated directly into the training loop as data augmentations in PyTorch (version 2.0). These operations, previously treated as fixed preprocessing steps, are *dynamically* applied during training. By embedding these augmentations into the training process, we simulate the variability and imperfections present in real-world diffraction patterns, allowing the model to better generalize across diverse experimental conditions. This approach emphasizes the importance of maintaining flexibility in the simulated data pipeline while aligning with the principles of self-supervised learning to enhance the model's robustness against natural perturbations in the input data. We call this model **CNNaug**.

Nonetheless, we use the same testing strategies as in the previous model (**CNNstatic**). For the **CNNaug** model, the classification accuracy of the crystal system (CS), extinction group (EG), and space group (SG) on the equivariance test set is 90%, 83%, and 81%, respectively. The accuracy on the invariance test set is significantly lower, however, with 45% for CS, 35% for EG, and 28% for SG. Similarly, the results on the experimental test set remain low, with classification accuracies of 23% for CS, 16% for EG, and 13% for SG.

The results of these experiments show similar trends compared to the supervised learning model, CNNstatic, with only slight improvements in mitigating overfitting. However, this experiment establishes the groundwork for the final self-supervised learning model proposed in this paper, paving the way for a more robust approach to addressing the challenges highlighted so far.

4.5. Self-Supervised Contrastive Representation Learning

In this Section, we introduce a methodology that incorporates the principles of representation learning within the previously described framework of self-supervised learning and finally present our model **DIFCON**. To achieve this, we modify the model architecture to include a representation-learning head (RH) positioned before the CS, EG, and SG classification heads. The RH is designed to optimize a contrastive learning objective, enabling the model to learn more robust and generalizable representations of the diffraction patterns. The fundamental idea behind using a contrastive learning [28,29] approach is to learn meaningful representations that model the previously discussed invariances and equivariances, by distinguishing between similar (positive) and dissimilar (negative) data samples. The central idea is to map similar inputs closer together in the learned feature space while pushing dissimilar inputs farther apart. For this application, as illustrated in Figure 3, a positive pair consists of two simulated diffraction patterns originating from the same crystal structure but different augmentation parameters, whereas a negative pair consists of diffraction patterns corresponding to different crystal structures.

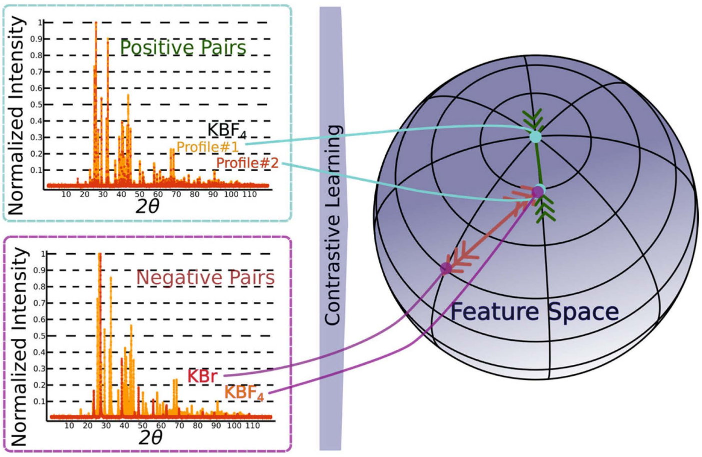

Figure 3. Visualizing the contrastive learning framework, with positive (KBF_4 vs. KBF_4) and negative (KBr vs. KBF_4) pairs.

The contrastive objective function encourages the model to focus on structural features being invariant to noise, sample variations, and experimental perturbations, while simultaneously maximizing the separability of different crystal structures. This design aligns with the overarching goal of achieving both equivariance to structural differences and invariance to experimental effects, as discussed in Section 3.

- By leveraging contrastive representation learning, the RH generates embeddings that capture essential features of the input diffraction patterns, which are subsequently passed to the downstream classification heads for CS, EG, and SG predictions. This approach not only strengthens the model's ability to generalize across simulated and real-world data but also facilitates more accurate indexing by learning a feature space that mirrors the underlying crystallographic distinctions. We consider two distinct contrastive learning approaches for **DIFCON**: *SimCLR* [28] and *Barlow Twins* [29]. Both approaches aim to learn robust representations but differ significantly in their objectives and optimization strategies, which we highlight below: *SimCLR* relies on a contrastive loss function called NT-Xent (Normalized Temperature-scaled Cross Entropy Loss). It uses positive pairs (augmented views of the same sample) and negative pairs (views of different samples) to define the loss. In the context of XRD, we adapt the *SimCLR* approach by using diffraction-specific augmentations, such as noise injection, random peak shifting, and impurity peak addition, to create positive pairs. Negative pairs are generated using diffraction patterns from different crystal structures.

- The *Barlow Twins* method, in contrast, eliminates the need for explicit negative pairs. It introduces a redundancy-reduction loss that aligns positive pairs while discouraging redundancy in the feature space. Specifically, the method aims to make the cross-correlation matrix of embeddings from positive pairs as close to the identity matrix as possible. By reducing redundancy, *Barlow Twins* ensures that each dimension of the learned representation captures unique information. A notable advantage of this method is its computational efficiency, as it does not rely on large batch sizes or negative samples. It does, however, require a much higher dimensional feature vector.

Using the *SimCLR* approach to train the RH, we observe a nice pattern when looking at the cosine similarities of the learned feature representation for positive and negative pairs. Figure 4 shows the cross-correlation matrix between the feature representations of 100 randomly selected crystal structures from the ICSD test phases, each of which was simulated using sampling strategies of the *invariance test set*, to produce eight different experimental effects—such as zero-shift, x- and y-axis noise, impurity peaks, and peak profile variation—and arranged consecutively. Ideally, this should follow a block diagonal structure. The figure compares this matrix to the corresponding matrix for **CNNaug** using cosine similarities of the internal features learned in its penultimate layer.

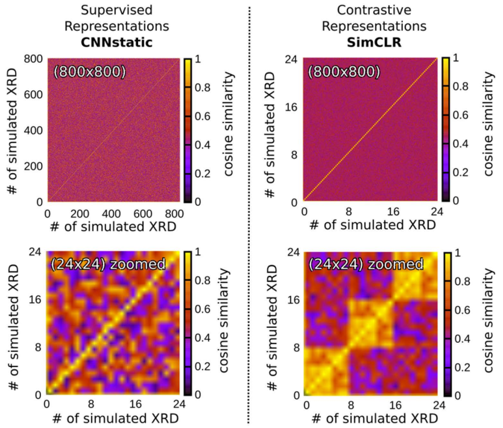

Figure 4. The cosine similarities (color bar) of the feature representations learnt by our supervised learning model **CNNstatic** (**left**) and the **SimCLR**-based contrastive learning model (**right**) for 100 randomly selected phases from a test set of phases. Each was simulated with 8 different experimental effects (like zero-shift, x- and y-axis noise, impurity peaks, and peak profile variation) and arranged consecutively.

This motivates our approach of using a contrastive learning objective to train the Representation Head (RH). In practice, however, training with the *SimCLR* approach requires an extremely large batch size to generate a sufficiently diverse set of negative pairs. This results in significantly longer training times for the RH to converge to an acceptable level of performance, particularly when trained jointly with the classification heads for the crystal system (CS), extinction group (EG), and space group (SG).

In contrast, we observed that the *Barlow Twins* approach, which does not rely on a large batch size of negative pairs, enables faster convergence of the RH and classification heads when trained end-to-end. Moreover, the *Barlow Twins* method demonstrates better classification performance on the invariance test set, likely because it can leverage more positive pairs within each training batch. Based on these observations, we adopted the *Barlow Twins* approach to train the final model presented in this paper, **DIFCON**.

The overall model architecture used by DIFCON is shown in Figure 5 on the left. The architecture is similar to that of CNNstatic and CNNaug concerning the 1D ResNet [26] encoder and the classification heads. The RH projects the output of the ResNet encoder onto a high-dimensional vector. The optimization process, shown in Figure 5 on the right, takes a batch of diffraction patterns and augments them in two different ways, passing each batch through the encoder and the RH separately. The output of the RH (z) for these two differently augmented batches is then used to compute a cross-correlation matrix (C) and is optimized to bring it closer to an identity matrix, using the following loss function $\mathcal{L}_{BT}$.

$$\mathcal{L}_{BT} = \sum_i (1 - C_{ii})^2 + \tau \cdot \sum_i \sum_{j \neq i} {C_{ij}}^2 \tag{6}$$

Here, τ is a positive constant, which is a hyperparameter that trades off the first and the second term of the loss function. The first term is responsible for the invariance between augmentations, and the second term is responsible for reducing the redundancy of the projected feature space. C is the cross-correlation matrix computed between the RH output of the differently augmented batches (z^A, z^B) along the batch dimension (b):

$$C_{ij} = \frac{\sum_b z^A_{b,i} \cdot z^B_{b,i}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \cdot \sqrt{\sum_b \left(z^B_{b,i}\right)^2}} \tag{7}$$

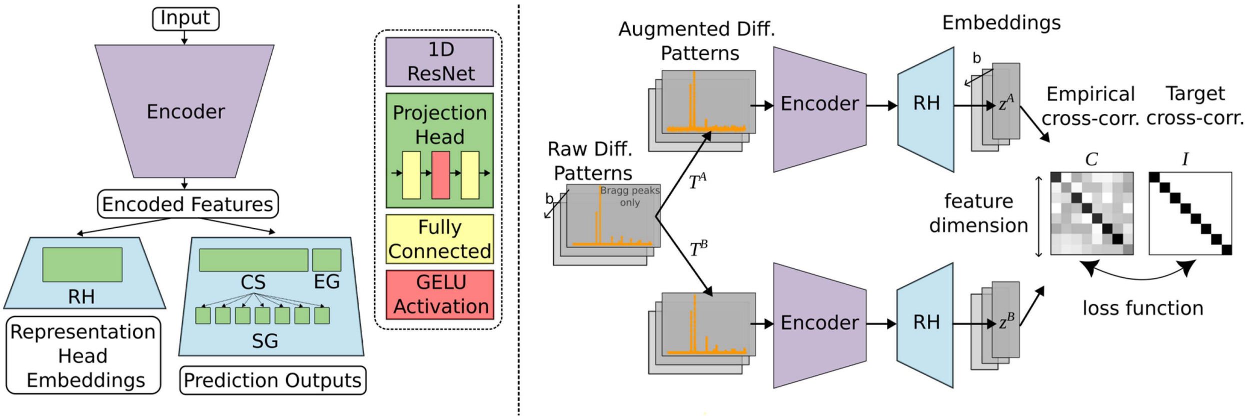

Figure 5. DIFCON model architecture (**left**) in which the encoder and the projection heads use a so-called GELU activation function [30]. DIFCON's optimization strategy (**right**) for learning representations, based on Barlow Twins [29]. A much larger feature dimension is scaled down for better visualization of the cross-correlation matrices.

After training the RH, the model is fine-tuned to the CS, EG, and SG prediction task using the same training strategy as CNNaug. For the equivariance test set, DIFCON achieves classification accuracies of 88%, 80%, and 78% for the CS, EG, and SG tasks, respectively. On the invariance test set, DIFCON shows a marked improvement, achieving classification accuracies of 79%, 71%, and 66% for CS, EG, and SG, respectively. These results highlight a significant improvement in the model's robustness and invariance to experimental effects simulated by our augmentation pipeline.

When tested on the experimental test set, DIFCON also shows notable advancements. The model correctly classifies the crystal system in 61 out of 82 instances (74%), which represents a significant improvement over both the earlier data-driven models and the results reported with NTREOR, too. Additionally, DIFCON achieves classification accuracies of

48% (39 out of 82) and 41% (34 out of 82) for the EG and SG tasks, respectively. All numerical results in this context are summarized in Table 2.

Table 2. Classification accuracies for all computational experiments. The equivariance and invariance test sets are simulated XRD patterns. NTREOR does not have an explicit equivariance test set; neither NTREOR nor HAND has an explicit invariance test set. The simulated (equivariance and invariance) test sets for CNNstatic, CNNaug, and DIFCON are precomputed and identical.

	Equivariance Test			**Invariance Test**			**Experimental Test**		
	CS	**EG**	**SG**	**CS**	**EG**	**SG**	**CS**	**EG**	**SG**
NTREOR	–	–	–	–	–	–	49%	–	–
HAND	94%	91%	87%	–	–	–	55%	32%	39%
CNNstatic	89%	82%	79%	40%	33%	24%	22%	15%	13%
CNNaug	90%	83%	81%	45%	35%	28%	23%	16%	13%
DIFCON	88%	80%	78%	79%	71%	66%	74%	48%	41%

5. Conclusions

This work attempts to describe the underlying bottlenecks for reliably using data-driven machine learning models needed for making accurate structural predictions from powder XRD patterns. Such ML models typically have a static inference time much lower than other algorithms like NTREOR and are theoretically more suitable in real-time environments. For instance, a single forward pass through the encoder and the prediction head of our biggest neural network model requires about 0.07 s on a local CPU node.

We show that training ML models using handcrafted features to predict specific structural properties like crystal systems, extinction groups, and space groups can achieve competitive performance when compared with well-known search-based algorithms like NTREOR. We observe that such models are quite robust to perturbations in the input. Such an approach, however, relies on manual human intervention and feature descriptions that are inherently specific to a particular task. For this reason, we explore using neural network models with entire diffraction patterns as inputs.

The lack of labeled experimentally measured XRD data is identified as the main bottleneck for training such ML models. Based on the sizeable but still numerically limited amount of data in terms of reported crystallographic information, that very information is used to generate a much larger (infinite in principle) amount of simulated data. These simulated data, upon incorporating so-called augmentations, eventually allow for learning self-supervised representations, by modeling the deviations from ideal diffraction pattern that arise due to experimental conditions, such as instrumental or sample effects. The relationship between experimental and simulated data includes a two-axis approach, one that varies due to experimental effects and one that is due to crystallographic (structural) differences.

This work addresses the challenge of learning meaningful representations from simulated diffraction patterns, specifically those invariant to noise or experimental effects and equivariant to structural changes in the crystal. The classification accuracies over the different test sets are used to indicate (a) how good the model is at learning meaningful representations from simulated diffraction patterns, and (b) how effective the signal processing operations (or augmentations) of the simulation pipeline are in modeling experimentally measured diffraction patterns. The relatively poor performance of real experimental data can be attributed to the problem of correctly modeling natural noise present in the data. This explains why models trained only with a supervised learning objective perform better when using handcrafted features as inputs rather than the entire diffraction pattern, as the burden of robustness lies on the user rather than the model. We managed to address this issue using our representation learning strategy. Using a contrastive learning objective, it is shown that the model can learn more useful representations, performing better than classically supervised learning objectives, for testing across both the invariance and the equivariance axis.

Our current approach is limited by the explicit types of experimental noise that we model and add when simulating our data. This could, however, be extended by better describing specific types of instrumental effects or even to neutron diffraction experiments. We are also limited by the fact that we focus only on diffraction patterns of primarily pure single-phase samples. We investigate ML models for prediction tasks (symmetry groups), where the models are trained to make one confident prediction, although it is important to note that such an approach is inherently flawed. An XRD measurement can lead to multiple candidate solutions, and we cannot expect to train an ML prediction model to achieve perfect results. Future work is likely to concentrate on generative [20] or exploratory neural network AI models, which can be designed to generate multiple candidate solutions (unlike prediction models that are trained to make one confident prediction) for solving more complicated tasks in the structure determination pipeline. The representation learning method proposed here can be scaled to fit such tasks. For instance, the learned feature representations can be used as feature embeddings, or as conditionals in probabilistic generative neural network models.

Author Contributions: Conceptualization, methodology, formal analysis, investigation, software, visualization, and writing: S.D.; investigation, experimental XRD data collection, test data curation and annotation: M.V.; conceptualization, investigation, validation, resources, supervision, funding acquisition, project administration, and editing: A.H.; project administration, validation, supervision, funding acquisition, and editing: R.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the consortium DAPHNE4NFDI in the context of the work of the NFDI e.V. The consortium is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) with project number 460248799.

Data Availability Statement: The ICSD structure data used in this work are proprietary and not provided here, but are commercially available. The experimentally measured XRD data and the codes for preparing and training the ML models presented in this paper are available from https://daphne.rwth-aachen.de/crystal-ai (accessed on 22 April 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Giacovazzo, C.; Monaco, H.L.; Artioli, G.; Viterbo, D.; Milanesio, M.; Gilli, G.; Gilli, P.; Zanotti, G.; Ferraris, G.; Catti, M. *Fundamentals of Crystallography*; Oxford University Press: Oxford, UK, 2011.
2. Giacovazzo, C. *Direct Phasing in Crystallography: Fundamentals and Applications*; Oxford University Press: Oxford, UK, 1998.
3. Brunger, A.T. Simulated annealing in crystallography. *Annu. Rev. Phys. Chem.* **1991**, *42*, 197–223. [CrossRef]
4. Kariuki, B.M.; Serrano-González, H.; Johnston, R.L.; Harris, K.D. The application of a genetic algorithm for solving crystal structures from powder diffraction data. *Chem. Phys. Lett.* **1997**, *280*, 189–195. [CrossRef]
5. Oszlányi, G.; Sütő, A. The Charge Flipping Algorithm. *Acta Crystallogr. A* **2007**, *64*, 123–134. [CrossRef]
6. Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; Song, D. Natural Adversarial Examples. *arXiv* **2019**, arXiv:1907.07174.
7. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
8. Altomare, A.; Giacovazzo, C.; Guagliardi, A.; Moliterni, A.G.G.; Rizzi, R.; Werner, P.-E. New Techniques for Indexing: N-TREOR in EXPO. *J. Appl. Cryst.* **2000**, *33*, 1180–1186. [CrossRef]
9. Gasparotto, P.; Barba, L.; Stadler, H.-C.; Assmann, G.; Mendonça, H.; Ashton, A.W.; Janousch, M.; Leonarski, F.; Béjar, B. *TORO Indexer*: A *PyTorch*-based indexing algorithm for kilohertz serial crystallography. *J. Appl. Crystallogr.* **2024**, *57*, 931–944. [CrossRef]

10. Suzuki, Y.; Hino, H.; Hawai, T.; Saito, K.; Kotsugi, M.; Ono, K. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. *Sci. Rep.* **2020**, *10*, 21790. [CrossRef]
11. Park, W.B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S.P.; Pyo, M.; Shin, N.; Sohn, K.-S. Classification of crystal structure using a convolutional neural network. *IUCrJ* **2017**, *4*, 486–494. [CrossRef]
12. Caglioti, G.; Paoletti, A.; Ricci, F. Choice of collimators for a crystal spectrometer for neutron diffraction. *Nucl. Instrum.* **1958**, *3*, 223–228. [CrossRef]
13. Lee, B.D.; Lee, J.-W.; Park, W.B.; Park, J.; Cho, M.-Y.; Singh, S.P.; Pyo, M.; Sohn, K.-S. Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction. *Adv. Intell. Syst.* **2022**, *4*, 2200042. [CrossRef]
14. Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N.T.P.; Ramasamy, S.; DeCost, B.L.; Tian, S.I.P.; Romano, G.; et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **2019**, *5*, 60. [CrossRef]
15. Salgado, J.E.; Lerman, S.; Du, Z.; Xu, C.; Abdolrahim, N. Automated classification of big X-ray diffraction data using deep learning models. *npj Comput. Mater.* **2023**, *9*, 214. [CrossRef]
16. Lolla, S.; Liang, H.; Kusne, A.G.; Takeuchi, I.; Ratcliff, W. A semi-supervised deep-learning approach for automatic crystal structure classification. *J. Appl. Crystallogr.* **2022**, *55*, 882–889. [CrossRef]
17. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *arXiv* **2014**, arXiv:1406.2661. [CrossRef]
18. Lange, J.; Komissarov, L.; Lang, R.; Enkelmann, D.D.; Anelli, A. Automatic Solid Form Classification in Pharmaceutical Drug Development. *arXiv* **2024**, arXiv:2411.03308.
19. Schulte, H.; Hoffmann, F.; Mikut, R. Siamese Netwroks for 1D Signal Identification. In Proceedings of the 30th Workshop Computational Intelligence, Berlin, Germany, 26–27 November 2020.
20. Lai, Q.; Xu, F.; Yao, L.; Gao, Z.; Liu, S.; Wang, H.; Lu, S.; He, D.; Wang, L.; Zhang, L.; et al. End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction. *Adv. Sci.* **2025**, *12*, 2410722. [CrossRef]
21. Sohl-Dickstein, J.; Weiss, E.A.; Maheswaranathan, N.; Ganguli, S. Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *arXiv* **2015**, arXiv:1503.03585.
22. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239.
23. Toby, B.H.; Von Dreele, R.B. GSAS-II: The Genesis of a Modern Open-Source All Purpose Crystallography Software Package. *J. Appl. Cryst.* **2013**, *46*, 544–549. [CrossRef]
24. O'Donnell, J.H.; Von Dreele, R.B.; Chan, M.K.Y.; Toby, B.H. A scripting interface for *GSAS-II*. *J. Appl. Crystallogr.* **2018**, *51*, 1244–1250. [CrossRef]
25. Altomare, A.; Cuocci, C.; Giacovazzo, C.; Moliterni, A.; Rizzi, R.; Corriero, N.; Falcicchio, A. *EXPO2013*: A kit of tools for phasing crystal structures from powder data. *J. Appl. Crystallogr.* **2013**, *46*, 1231–1235. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. McKay, M.D.; Beckman, R.J.; Conover, W.J. Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* **1979**, *21*, 239–245. [CrossRef]
28. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv* **2020**, arXiv:2002.05709.
29. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *Proc. Mach. Learn. Res.* **2021**, *139*, 12310–12320.
30. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (GELUS). *arXiv* **2016**, arXiv:1606.08415.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.