Application of Machine Learning to XRD Phase Identification

Jan Schützke[1,3], Brian Jones[2], Nathan Henderson[2], S. Nick Rodesney[2],
Alexander Benedix[3], Karsten Knorr[3] , Ralf Mikut[3], Markus Reischl[3]
[1] Bruker AXS GmbH, Karlsruhe, Germany
[2] Bruker AXS, Madison, WI - US
[3] Institute for Automation and Applied Informatics, KIT, Karlsruhe, Germany
Contact: Jan.Schuetzke@Bruker.com, Brian.Jones@Bruker.com

69th Annual Conference on Applications of X-ray Analysis (Denver X-ray Conference)

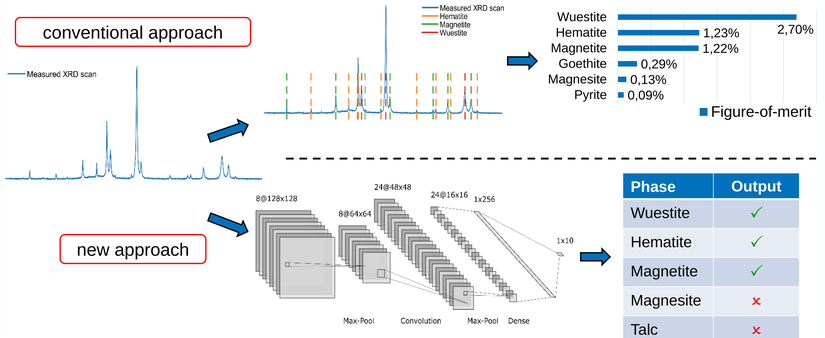
Identification of crystalline phases in mixtures is a frequently performed task in powder XRD. It mostly involves software for searching databases of known compounds, and matching lists of d-spacings and related intensities to the measured (or reduced) data. Figures-of-merit are usually taken as numerical indicators for the probability of the individual phase assignments. However, some expertise of the skilled user is still required for a "manual" validation of the results. This is time-consuming and error prone. Most automated search/match procedures that apply some iterative procedure of the above aim at making the validation step redundant but have failed to be generally reliable to this point.

In recent years, deep learning models established their status as a state-of-the-art approach for automated image analysis, such as detecting cars and pedestrians in a street scene. In analogy, deep learning models were applied here for automated phase identification from one-dimensional XRD data. We used phases and mixtures from the Bruker AXS iron ore and cement application packages that are in wide commercial use and contain 28 and 76 phases, respectively. Several models have been tested, which learn the peculiarities of XRD data to support the automated phase identification process. A framework for the efficient generation of hundreds of thousands of simulated scans has been developed, since real measured and labeled scans are only scarcely available, and deep learning approaches require an extensive dataset to learn a general representation. This learning step considered not only varying phase presence and concentrations but noise, background variation, peak broadening as well as variance in lattice parameters.

The trained networks achieve an accuracy of close to 100 % for synthetic mixtures of both application packages while analyzing hundreds of scans in under a second, thus outperforming the experts in speed without sacrificing an accurate prediction. Additionally, the models have been tested on measured XRD scans to confirm the results.

Concept: Automatic Phase Identification

Traditional phase identification approaches rely on matching d-spacings for observed reflections and calculating a figure-of-merit (FOM). We wanted to avoid hard-coding such analysis criteria and methods and instead use a trainable, machine learning algorithm to provide a simple binary prediction for the presence of a given phase (i.e., 1 = present, 0 = absent).

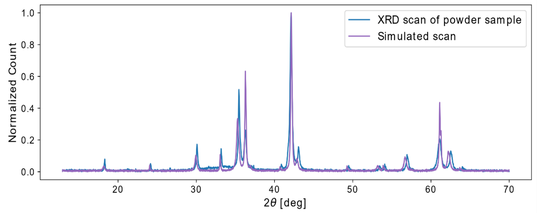


Phase	Output
Wuestite	✓
Hematite	✓
Magnetite	✓
Magnesite	×
Talc	×

Generation of Training Data Sets

Recent advances in computational power provide the capability for automated analysis; however, machine learning algorithms still require a suitable training set comprised of thousands of scans. This represents a significant challenge from an experimental standpoint.

To address this, we generated training datasets with DIFFRAC.TOPAS, which offers a scriptable interface for simulating scans. Synthetic scans were created for mixtures with varying weight percentages as well as commonly observed contributions like background, Gaussian noise, and air-scatter. Two applications packages with a restricted number of candidate phases were used: iron ore (28 phases) and cement (76 phases). From these, we generated 500,000 synthetic scans for training and an additional 100,000 scans for validation. Each training scan was supplemented with a list of all present and absent phases.

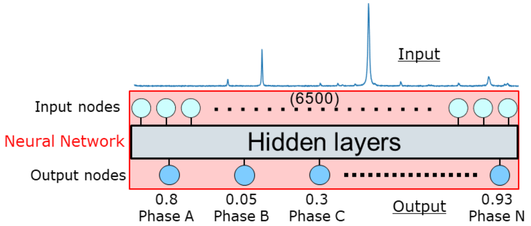


Use of Neural Networks and Results

Artificial neural networks were used towards automatic phase identification. Raw scans with identical scan ranges (5-70 degrees 2-Theta) and step widths (0.01 degrees) were provided as inputs. The network outputs a certainty score between 0-1 for the presence of each phase.

Training of the neural network is accomplished by optimizing the cross-entropy of the present/absent binary. The raw training scans propagate through the network and create an output. Binary cross-entropy is calculated by comparing the target and actual outputs. The weighting schemes (within the hidden layers of the neural network) are adapted by back-propagation, and the network is trained by annealing the actual output of the neural network to the target output for each sample. Ultimately, the network will settle on a good generalization that fits all scans provided in the training set.

After the network was trained, validation scans were processed through the algorithm to test the capabilities for automatic phase identification. During application, phases with an output value greater than 0.5 count as predicted present, while smaller than 0.5 is interpreted as absent. Results are shown in the table below.



Application Package	Number of Candidates	Accuracy	F1-Score
Iron Ore	28	99.99 %	99.54 %
Cement	76	99.99 %	98.60 %
Combination	345	99.99 %	91.30 %

Summary

- Automatic phase identification via trained neural networks is possible with an accuracy of nearly 100 % for the tested application packages (iron ore and cement)
- The number of false positives increases with larger numbers of candidates
- The trained network operates close to human performance levels while analyzing hundreds of scans per second
- Future work will target a model for packages with a larger number of phases or an approach that is transferable between applications

X-ray Diffraction