Machine Learning-based Crystal Structure Prediction for X-Ray Microdiffraction

Yuta Suzuki[1,3], Hideitsu Hino[2], Yasuo Takeichi[3], Takafumi Hawai[3], Masato Kotsugi[1], Kanta Ono[3,*]

1. Tokyo University of Science, Tokyo, Japan., 2. The Institute of Statistical Mathematics, Tokyo, Japan., 3. High Energy Accelerator Research Organization, Ibaraki, Japan.

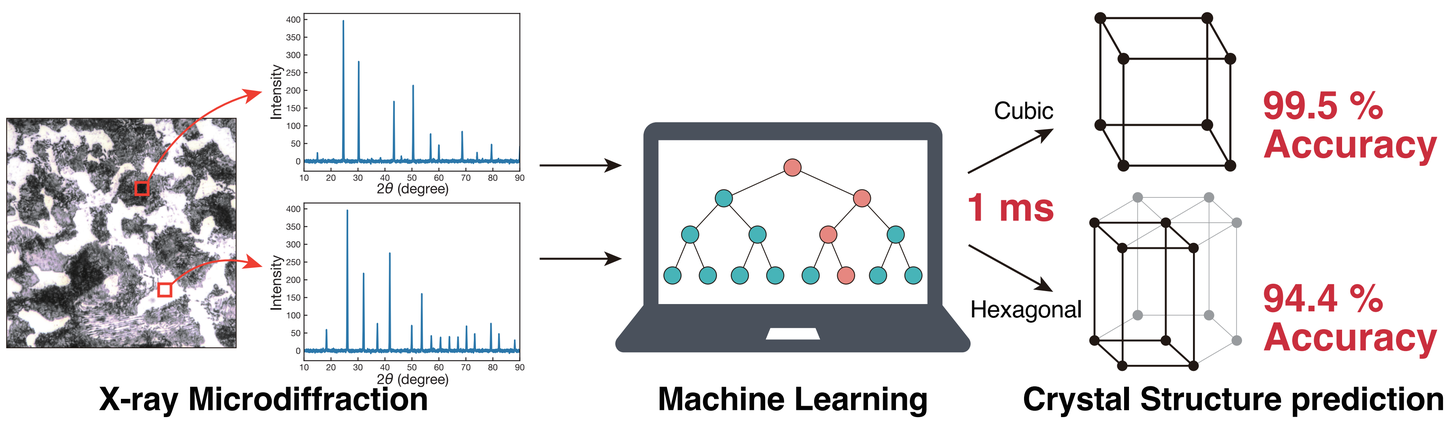
Introduction

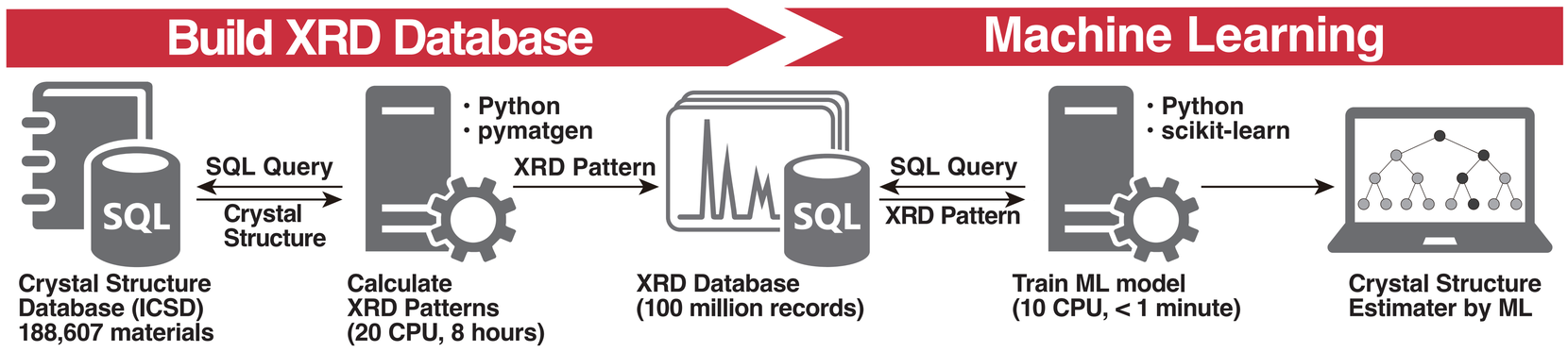
- **Materials Informatics (MI)**
  Acceleration of materials discovery and obtaining knowledge by statistical learning of materials data.
- **Key of accelerated materials discovery**
  1. Combinatorial systhesys of materials
  2. High-throughput experiments
  3. On-the-fly data analysis
- **X-ray Diffraction (XRD)** is one of most important technique for materials characterization. But, **the conventional analysis is performed manually**, it could be a **bottleneck** of materials discovery workflow.
- Estimation of a crystal structure from XRD pattern is difficult, since try-and-errors are required.
  If we build machine learning(ML) model to predict the crystal structure from a XRD pattern, the data analysis will be automated and accelererated, and XRD will become more powerful method.

Objectives

- **Can we automatically estimate crystal structures (crystal system, space group) from XRD patterns with machine learning?**
- **Can we obtain knowledges from materials data with machine learning?**
- We are aiming to accelererate materials discovery by the integtaration of the high-throughput materials measurement and on-the-fly data analysis.

Our strategy and Methods





- International Crystal Structure Database (ICSD) was used for the data source. It contains 188,607 materials data. Errors of data were eliminated, 169,390 data were used.
- XRD patterns were calculated using pymatgen [1] middleware. The wavelength and 2θ range were set to CuKa radiation (1.5418 Å) and 0˚-90˚, respectively.
- The calculated XRD data is stored in Relational Database Management System(RDBMS).
- We considerd the descriptor for the XRD patterns, we chose the 2θ positions of first ten peaks and the number of all peaks within 0˚-90˚.
- Random Forest (RF) [2] was employed for the machine learning algorithm.
- RF is flexible and able to express complex nonlinear functions. It offers stable estimation result, fast processing speed. From these favorable characteristics, we applied RF to this research.
- The hyper parameters were optimized with grid search. The number of trees and max feature ratio were set to 500, 0.6, respectively.
- The generalization performance is estimated by 10-fold cross varidation.

Results and Discussion

Crystal System	Accuracy (%)
Cubic	99.49
Tetragonal	93.75
Hexagonal	94.38
Monoclinic	87.70
Trigonal	90.58
Orthorombic	90.21
Triclinic	47.18
All	91.88

Table.1 The accuracy of crystal structure estimation.

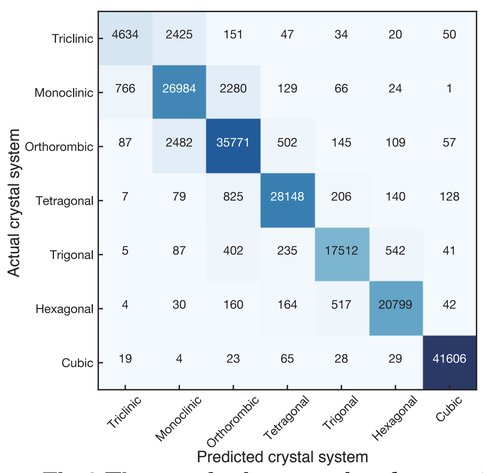


Fig.1 The confusion matrix of crystal structure estimation.

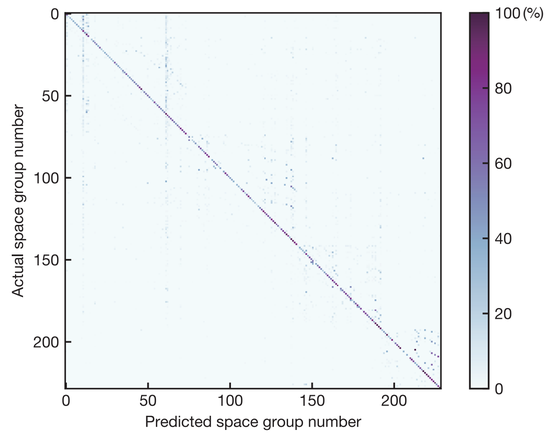
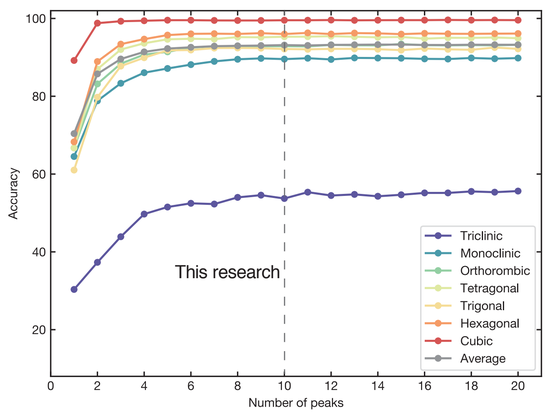


Fig.2 The confusion matrix of space group estimation. The accuracy is normalized in percent.



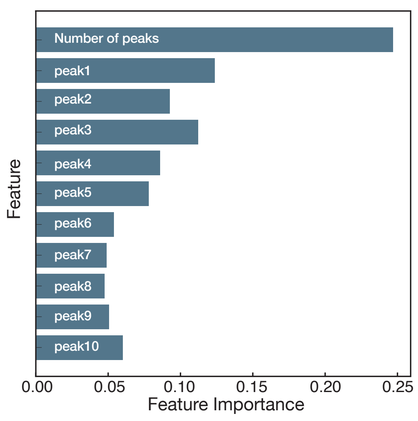


Fig.3 (Left) The effect of number of XRD peaks for estimation accuracy of crystal system.

Fig.4 (Right) The feature importance for crystal system estimation, calculated by RF.

- **The prediction accuracy of crystal system (seven classes) was 91.88 % (Tb. 1, Fig. 1).**
- **99.49 % accuracy for cubic system was obtained and the estimation took less than 1 ms per XRD pattern.**
- **The prediction accuracy of space group (230 classes) was 80.61 % (Fig. 2).**
- Our result suggests that **8 XRD peaks are sufficient for crystal system prediction, and especially in case of a cubic system, it is 2** (Fig.3).
- The trained RF model provides the importance of each input feature. **It revealed that the number of peaks and several peaks of lower 2θ are important to predict crystal system.**
- The relatively poor prediction performance for triclinic system was caused by the insufficiency of data. The structure is complicated, has six degrees of freedom. ICSD contains just 7300 triclinic materials (4 %); it should not be enough to train ML model as well. Data augmentation might help this problem.
- The prediction accuracies are similar to the recently reported deep learning results, that are 94.99 % and 81.14 % for crystal system and space group, respectively [3].

Summary

- **We built XRD database for 188,607 materials.**
- **We applied machine learning for estimation of crystal system and space group from XRD patterns. The prediction accuracy was 91.88 %, 99.49 %, 80.61 % for crystal system, cubic system, and space group, respectively.**
- **We are now working on development of ML-based automated XRD peak indexing and lattice constant estimation as well.**
- **We obtained knowledge of the influence of input feature effect on prediction accuracy, with combining the built XRD database and ML.**
- **New knowledge of materials science will be achieved by such data-driven research.**

Resources

- Please visit our GitHub and website for download this poster and contact us.

github.com/resnant/XRM2018

resnant.github.io

References

1. Ong, S. P. et al. Comput. Mater. Sci. **68**, 314–319 (2013).
2. Breiman, L. Random Forests. Mach. Learn. **45**, 5–32 (2001).
3. Park, W. B. et al. IUCrJ **4**, 1–9 (2017).