Abstract for an Invited Paper
for the MAR08 Meeting of
The American Physical Society

Materials Informatics: Using machine learning techniques with large amounts of ab-initio computed or experimental data

GERBRAND CEDER, Massachusetts Institute of Technology

Machine learning techniques can be applied to large amounts experimental or computed materials data in order to identify the underlying factors that determine a target property. While the use of experimental data is complicated by the fact that it is mostly non-standardized in property or structure databases, experimental data still tends to be richer in information than computed data. One problem that can be addressed with machine learning techniques is the prediction of structure. By using structure prototype as a mathematical descriptor, and constructing its correlation in chemical spaces through machine learning techniques, it is possible to create a highly effective structure prediction method. Previously, we demonstrated that by simply applying maximum entropy ideas to a large experimental structure database of binary metals, it was possible to suggest a short list of candidate structures for new compounds which contains the proper ground state with very high probability [Ref]. This list of probable structures can then be computed with ab initio energy methods. We have now extended this method to multi-component and non-metal systems by prototyping the $\approx$ 100,000 structure records in the International Crytallographic Structure Database, and a similar accuracy of prediction is achieved in these high component spaces. We believe that such a machine learning approach solves the crystal structure prediction for many practical purposes. Machine learning techniques can also be used to point at likely errors in experimental structure databases and I will give some examples of this. In the long-term computed data is more likely to form the input for machine learning techniques as it is well defined and obtained under controlled conditions. Using high-throughput ab-initio computing techniques we have determined the structure and energy for several thousand compounds and have begun to data mine this information for property models relevant to energy generation and storage.