Development of an algorithm to efficiently extract chemistry descriptors in organically templated metal oxides

Xiwen Jia, Alexander J. Norquist, Joshua Schrier
Department of Chemistry, Haverford College
370 Lancaster Avenue, Haverford PA

Organically templated metal oxides are a diverse class of compounds that have attracted sustained interest over the years due to their high degree of compositional and structural diversity and technologically desirable physical properties.[1] Even though machine-learning algorithms trained on failed hydrothermal syntheses can predict whether crystals will form or not,[2] predicting structures remains a challenge. However, we hypothesize that descriptive features of the structures---such as layer-dimensionality and atom coordination number---can be predicted using properties of the reagents, stoichiometry and reaction condition.

In this project, an algorithm was written using the CSD python API to efficiently extract structural descriptors of organically templated metal oxides directly from the Crystallographic Information Files (CIFs) so that we can learn them using a machine-learning approach. Organically templated vanadium tellurites were chosen as a system to begin with because their known structural variety would be helpful in developing and testing the algorithm. All 21 reported vanadium tellurites structures were extracted as CIFs from the Cambridge Crystallographic Data Center (CCDC). The chemistry descriptors of interest in the secondary building units (SBUs) of vanadium tellurites are the following: vanadium-vanadium minimum atomic distance, tellurium-tellurium minimum atomic distance, vanadium coordination number, tellurium coordination number, the presence of vanadium-oxygen dimers, the presence of tellurium-oxygen dimers and layer-dimensionality. All these descriptors contain information about the crystal structures of vanadium tellurites, as indicated by scatterplots, k-means clustering, histograms, principal component analysis and decision tree. In order to check the accuracy of my extraction, all SBUs of the 21 vanadium tellurites structures were plotted in VESTA (Visualization for Electronic and STructural Analysis). Their 7 structural descriptors were manually extracted and recorded. The developed algorithm can successfully extract all 7 chemistry descriptors of the SBUs in 21 organically templated vanadium tellurites structures, with an accuracy rate of 100%. In addition, the short running time of the code makes it a promising algorithm for future usage. The running time of the dimensionality section ranges from 16 seconds to 4 minutes while that of other chemistry descriptors sections is within seconds.

In the next stage of this project, the software will be generalized for vanadium compounds, such as vanadium selenites, vanadium borates and vanadium sulfites, and even extended to organohalide perovskites. In addition, machine-learning models will be developed to predict structural outcomes using the reagent properties, stoichiometry, and reaction condition.

1. Cheetham, A. K.; Rao, C. N. R.; Feller, R. K.Chem. Commun. **2006**, 4780–4795
2. Raccuglia, P. et. al. Machine-learning-assisted materials discovery using failed experiments. Nature. **2016**, 533, 73–76