University of Utah
UNDERGRADUATE RESEARCH JOURNAL

IS CHEMICAL DOMAIN KNOWLEDGE EVEN NECESSARY WHEN MACHINE LEARNING MATERIAL PROPERTIES?

Kaai Kauwe, Ryan Murdock, Dr. Taylor D. Sparks
Department of Materials Science and Engineering

ABSTRACT

The process of predicting material properties using machine learning often involves engineering a description of the materials being assessed, which is then used as input to various models for regression or classification. This description commonly comes in the form of a list of easily-measured characteristics of elements in the material. For instance, one might include the boiling point, atomic radius, and electronegativity of each element within a particular chemical formula. Creation of such composition-based feature vectors (CBFVs) can be time-intensive and requires significant domain knowledge. The advent of learned elemental embeddings and elemental encodings created with no explicit knowledge of chemistry challenges the practice of utilizing and creating CBFVs. Further, it raises questions concerning the necessity of domain knowledge for the practice of material informatics. This work assesses the efficacy of CBFVs and chemistry-free representations given different predicted properties and dataset sizes in order to compare these two approaches. We find that the simple one-hot encoding of elements performs competitively with other representations under some circumstances. For instance, preliminary results indicate that using one-hot encoding with a simple model and a large dataset may reduce the mean absolute error (MAE) of predicting shear modulus by 2.5% when compared to Mat2Vec, a learned embedding. Further, these preliminary results indicate a potential 9.5% decrease in MAE with one-hot on predicting formation energy when compared to Magpie, a CBFV.