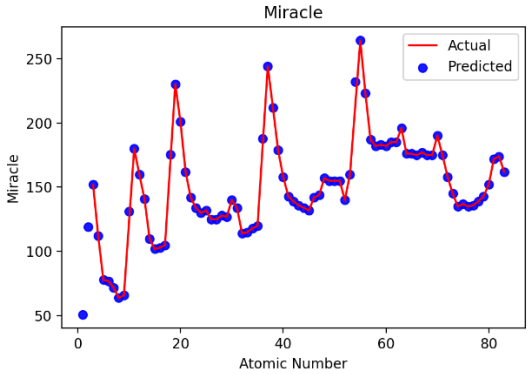
Machine-learned features to solve crystal structure classification problems

Sangjoon Lee[1], Anton O. Oliynyk[2]

[1]*Department of Chemical Engineering, The Cooper Union for the Advancement of Science and Art, New York, NY 10003*

[2]*Department of Chemistry and Biochemistry, Manhattan College, Riverdale, NY 10471*



Materials informatics uses data-driven approaches for the study and discovery of materials properties. Data availability and features are two crucial components in generating reliable and accurate machine-learning models. While general data can be acquired through public and commercial sources, features must be tailored for a specific application. Common featurizers, such as Atom2Vec, Jarvis, mat2vec, and Magpie, suitable for generic chemical problems, may not be ideal for solid state materials. Here, we have identified our in-lab compiled feature list which outperforms the aforementioned featurizers for solid state materials with 50 to 1,000 training data, a common dataset size in materials informatics. We applied Gaussian process regression (GPR) to predict missing values. A complete dataset in the feature list allows researchers to apply our features to any methods that require complete x-block. To validate our updated feature list, a classical crystallographic problem of classifying structure type was solved. Similarly to the radius ratio rule (Linus Pauling, 1929) and structure maps (Villars, 1983; Pettifor 1984), we demonstrate how binary equiatomic solid structures (CsCl-, NaCl-, and ZnS-types) can be successfully segregated with machine learning using our proposed feature list.